The Pennsylvania State University

The Graduate School

College of Education

DIAGNOSTIC ACCURACY OF THE CULTURE-LANGUAGE INTERPRETIVE MATRIX WITH THE WJ-III-NU: A COMPARISON OF SPANISH-SPEAKING ENGLISH LANGUAGE LEARNERS AND MONOLINGUAL ENGLISH-SPEAKING STUDENTS

A Dissertation in

School Psychology

by

Erin Lynne Meyer

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2013

The dissertation of Erin L. Meyer was reviewed and approved* by the following:

Beverly J. Vandiver
Associate Professor of Education
Dissertation Advisor
Chair of Committee

James C. DiPerna
Associate Professor of Education
Professor in Charge of the School Psychology Program

Keith B. Wilson
Professor of Education

Janet van Hell
Professor of Psychology and Linguistics
Director of the Linguistics Program
Professor of Language Development (Radboud U. Nijmegen, the Netherlands)

Kathleen J. Bieschke
Professor of Education
Department Head of Educational Psychology, Counseling, and Special Education

*Signatures are on file in the Graduate School

ABSTRACT

The purpose of this study was to examine the appropriateness of using the Culture-Language Interpretive Matrix (C-LIM; Flanagan, Ortiz, and Alfonso, 2007) with the Woodcock-Johnson – Third Edition, Normative Update, Tests of Cognitive Abilities (WJ-III-NU; Woodcock, McGrew, Schrank, & Mather, 2007) when assessing school-aged English language learners (ELLs). An archival sample of 78 referred Spanish-speaking ELLs and 156 referred monolingual English-speaking students was gathered and matched based on age and gender. The C-LIM was interpreted using three levels of criteria: (a) most stringent, based on a nine cell decline, (b) moderately stringent, based on a five level decline, and (c) least stringent, based on a three cell decline. Within each criterion, three levels of decline were examined: (a) cultural loading, (b) linguistic demand, and (c) combined cultural loading and linguistic demand. Primary analyses were diagnostic accuracy statistics and receiver operating curve (ROC) analysis. Findings indicated that although the C-LIM interpretation generally identified MESs correctly based on not following a pattern of decline, it did not accurately identify ELLs based on following a pattern of decline. Results of ROC analyses revealed that all types of C-LIM interpretation investigated resulted in similar diagnostic decisions for ELLs and MESs no matter which cutoff score was used to define a declining pattern. Analyses conducted on ELLs and MESs who were not diagnosed with specific learning disabilities produced similar results. The influences of acculturation and language proficiency were also examined relative to C-LIM decisions. Based on these findings, use of the C-LIM in practice is not recommended at this time. Recommendations are offered for research and practice involving the assessment of English language learners.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express gratitude to all those who have supported me during graduate school and the completion of my dissertation. First, I would like to thank my dissertation advisor and committee chair, Dr. Beverly Vandiver, for your time, knowledge, and diligence in guiding me through the dissertation process. I would also like to thank the other members of my doctoral committee, Drs. James DiPerna, Janet van Hell, and Keith Wilson, for your encouragement and expertise.

Next, I would like to thank Dr. Heather Applegate for your support throughout the completion of my dissertation. Your ability to ask the hard questions was particularly valuable. Special thanks go to Dr. Sonya Lanier, my internship supervisor, for always encouraging me whenever I needed it. I would also like to thank the research office and all of the psychologists at Loudoun County Public Schools for providing me the necessary data to complete this project.

To my school psychology cohort, especially Anne, Melissa, and Miranda, thank you for being such great friends. I would not have made it through grad school without you! To Sarah and Chad, I am so grateful for your encouragement and friendship. I also want to thank you for being so generous with your time and your house during my trips to State College.

To my family, thank you for believing in me, always being there to listen, and knowing that I would eventually finish. Finally, I would like extend my deepest gratitude to my husband, Dave. Words cannot adequately express my appreciation for your support. I truly would not have finished this dissertation without your patience, encouragement, and love.

This work is dedicated to my brother, Stefan James Bachmann. You have always been there for me, even after I could no longer see your smile, hear your laughter, and feel your bear hugs. I know that you are celebrating this accomplishment with me in spirit.

INTRODUCTION

According to recent census data (Humes, Jones & Ramirez, 2011), the U.S. population grew almost 10 percent in the past decade and more than half of that growth was due to an increase in the Hispanic population. In fact, non-Hispanic Whites have showed the slowest rate of growth over the same time span in comparison to all other racial and ethnic groups represented in the census (Humes et al., 2011). The increase in diversity is also evident in U.S. schools. A report from the U.S. Department of Education (Aud et al., 2011) indicated that racial and ethnic minority enrollment increased from 32 percent of the U.S. student population in 1989 to 45 percent in 2009. Between 1980 and 2009, the number of school-age children (ages 5-17) who spoke a language other than English at home more than doubled, from 4.7 to 11.2 million (Aud et al., 2011). In other words, as of 2009, 21 percent of the school-age population in the U.S., or one in five students, spoke a language other than English at home. Furthermore, five percent of school-age children spoke English with difficulty (Aud et al., 2011). These are just a few of the indicators that underscore the growth of diversity in U.S. schools, as well as the distinct challenge educators currently face in meeting the needs of a changing population.

Concerns about addressing the needs of culturally and linguistically diverse (CLD) students in the educational system are not new. The term CLD is used to describe individuals whose first language is not English and whose background and experiences differ from mainstream individuals born in the U.S. (Rhodes, Ochoa, & Ortiz, 2005). Educators typically refer to CLD individuals as English language learners (ELLs) and the term ELL will be used from here on. The free and appropriate education of ELLs has been the focus of court cases (*Diana v. California State Board of Education*, 1970; *Lau v. Nichols*, 1974), educational laws (Americans with Disabilities Act of 1990; Education for All Handicapped Children Act of 1975),

and professional codes of ethics for many years (*Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations*, American Psychological Association, 1990; *Principles for Professional Ethics*, National Association of School Psychologists, 2010; Rogers et al., 1999; *Standards for Educational and Psychological Testing*, American Educational Research Association, APA, & National Council on Measurement in Education, 1999).

Despite the actions of government and professional agencies to provide a free and appropriate education for all students, research indicates that ELLs are disproportionately represented in special education programs (Artiles & Trent, 1994; Artiles, Rueda, Salazar, & Higareda, 2005; Rueda & Windmueller, 2006). Disproportionate representation means unequal representation of students in special education based on characteristics such as race, ethnicity, and ELL status (Rhodes, Ochoa, & Ortiz, 2005). In recent years, disproportional representation of racial, ethnic, and linguistic minority groups in special education has been reframed in terms of addressing contributing factors (e.g., the referral, evaluation, and identification process for special education) instead of focusing entirely on measuring and reducing disproportionality (Coutinho & Oswald, 2006). Actions taken to reform instructional approaches and referral processes have shown promise (Gravois & Rosenfield, 2006; Green, McIntosh, Cook-Morales, & Robinson-Zañartu, 2005), but for ELLs who are referred for an evaluation to determine the presence of a disability, concerns regarding the assessment process for special education remain.

School Psychologists' Assessment Practices with English Language Learners

Psychoeducational evaluations for ELLs have markedly evolved since the 1970s, when court decisions and laws (i.e., *Diana v. California State Board of Education,* 1970; Education for All Handicapped Children Act, 1975) were first enacted requiring assessment in a student's first

language if at all possible. In a survey on assessment of ELLs, Bainter and Tollefson (2003) suggest that psychologists are in agreement regarding what is considered acceptable assessment practices. Two such practices are using bilingual psychologists for evaluations of ELLs who are more proficient in their first language, or testing in English when an ELLs' language proficiency in English is stronger than their proficiency in another language. However, the agreed upon assessment practices are often not feasible because many psychologists are not bilingual. Furthermore, those who are proficient in specific languages other than English could be responsible for evaluating all non-English speaking students, regardless of the language spoken. Bainter and Tollefson also found that psychologists differed in their opinions of other assessment options, such as the use of nonverbal tests, foreign-normed measures, and administration of tests in English when a student has higher proficiency in another language. Taking steps to account for language proficiency may allow psychologists to more clearly delineate its influence on students' cognitive test results. However, other factors may influence ELLs' cognitive scores (e.g., level of acculturation), and addressing language proficiency alone may not necessarily result in a valid standardized cognitive assessment. The results of Bainter and Tollefson's study highlight the need for training on appropriate assessment practices for ELLs.

Current "Best Practices in Nondiscriminatory Assessment" (Ortiz, 2008) suggest the use of a framework that encompasses not only actions related to special education evaluations, but also prereferral activities (Ortiz, 2008). The framework involves an overall focus on (a) intervention, (b) the use of authentic and alternative assessments, (c) evaluation of the learning environment and opportunities for learning, (d) assessment of language proficiency and language dominance (particularly in the event of formal testing), (e) evaluation of cultural and linguistic factors relevant to the student's education, (f) hypothesis development and testing, (g) attempts

to reduce bias in traditional assessment practices, and (h) support of conclusions based upon the convergence of data from multiple sources.

In addition to current guidelines and research within the field of school psychology, relevant findings in related fields, such as neuropsychology and applied linguistics, also offer valuable insight into the cognitive functioning and language skills of ELLs. For example, consistent differences in test performance on certain cognitive tasks between bilingual[1] and monolingual individuals have been found across the lifespan (Bialystok & Craik, 2010; Bialystok, Craik, & Luk, 2012). Bilingual individuals consistently perform better on executive function tasks (e.g., Stroop task), while monolingual individuals consistently perform better on verbal tasks (Bialystok, 2011a, 2011b). It is also important to be aware of aspects related to the student's first language other than level of proficiency, such as the level of maintenance of the first language (or level of shift to English) by the student's parents and family in their home (Bayley & Bonnici, 2009); maintenance of a first language reflects multiple characteristics, including family culture and adherence to old and new community and/or societal norms. These aspects of an ELL's cognitive functioning are important to consider, but could also be easily overlooked if a checklist approach toward ELL evaluations is adopted.

Utilization of multiple sources of information within a broad framework towards nondiscriminatory assessment may help to address the needs of many ELLs with academic difficulties; however, those who experience significant learning difficulties may still be referred for special education evaluations. Bilingual assessment is optimal when evaluating ELLs, but may not be possible given the small percentage of psychologists who are bilingual, the numerous languages present in the student population, and the dearth of bilingual assessment measures

[1] For the purposes of this study, bilingual is defined as one's ability to effectively use two languages. Monolingual is defined as one's ability to use one language. There are many possible definitions for these terms (Skutnabb-Kangas, 2007).

with appropriate norms. Nonverbal assessment is also typically used for evaluations of ELLs, but a limited number of specific cognitive abilities can be assessed through nonverbal means. For example, knowledge tasks typically require verbal communication and are not included in nonverbal intelligence measures. Therefore, for ELLs referred for special education evaluations, there is still a need to address weaknesses of the standardized cognitive instruments used. The cultural loading and linguistic demand of currently available cognitive assessments must be accounted for to ensure legally and ethically responsible evaluations of ELLs (Ortiz, 2008).

Cultural Bias in Standardized Assessments

A test is considered biased if the psychometric characteristics, including concurrent and predictive validity, differ across racial or ethnic groups (Brown, Reynolds, & Whittaker, 1999). A substantial amount of psychometric research (Brown et al., 1999; Reynolds, 2000; Sandoval et al., 1998, Valdés & Figueroa, 1994) has found no evidence of test bias in standardized intelligence measures. However, test bias does not take into account the developmental process of acculturation. On the other hand, *cultural loading*, a term used by Flanagan et al. (2007), is considered a fluid construct that "represents the degree to which a given test requires specific knowledge of or experience with mainstream U.S. culture" (p. 170). Thus, cognitive tests may not be culturally *biased*, but this fact does not mean they are not culturally *loaded*.

Cognitive measures (e.g., Woodcock-Johnson – Third Edition, Normative Update, Tests of Cognitive Abilities [WJ-III-NU; Woodcock, McGrew, Schrank, & Mather, 2007]) typically used in psychological evaluations have been standardized and normed on monolingual, English-speaking individuals born in the U.S. The resulting norms inherently reflect the experiences and subsequent knowledge of this population. In addition, the content of the tests is similarly influenced by the cultural background of the test developers. Test developers decide what items

to include, where to place them, and what responses are considered acceptable. These aspects of test development are ultimately influenced by the beliefs, attitudes, and experiences of the test developers. Therefore, cognitive measures are influenced by the culture in which they are developed and from which they are normed (Flanagan et al., 2007).

When interpreting the results of cognitive assessments, practitioners typically follow the "assumption of comparability" (Salvia & Ysseldyke, 1991): students to whom the test is administered have similar characteristics to those on whom the test was standardized and normed. If a student's acculturation level differs from those of the group on which the test was developed, the assumption of comparability has been violated. For example, when students' are less acculturated than the norm group (as may be the case for a student who has lived outside the United States), it would be inappropriate to interpret the results of the cognitive assessment as directly indicative of the students' current or future academic performance. Acculturation is "the social and psychological exchanges that take place when there is continuous contact and interaction between individuals from different cultures" (Cabassa, 2003, p.127). It is a developmental process and as individuals become accustomed to a new culture, the knowledge gained could improve their performance on culturally loaded tests (Flanagan et al., 2007).

Cognitive tests differ in the amount of culture-specific content included and tasks that are more novel in nature (e.g., Symbolic Memory on the Universal Nonverbal Intelligence Test, UNIT; Bracken & McCallum, 1998) may result in scores that are less influenced by level of acculturation (Jensen, 1974; Valdés & Figueroa, 1994). Thus, scores on tests that are less culturally loaded (i.e., more process oriented and novel in nature) may provide more accurate estimates of cognitive ability for individuals from culturally diverse backgrounds as a high level of acculturation is not necessary to grasp the intent of the task (Flanagan et al., 2007). In

summary, test bias may occur as a result of a school psychologist administering a cognitive test, which is culturally bound, to an ELL student despite the lack of standardization evidence to support its use.

Linguistic Bias in Standardized Assessments

Standardized tests are not classified as linguistically *biased* based upon psychometric characteristics, but they are, in various degrees, linguistically *demanding* (Flanagan et al., 2007). Linguistic demand is related to the level of English language proficiency of an examinee relative to same-age, English-only speaking peers (Flanagan et al., 2007). Tests that have long verbal directions and require elaborate verbal responses are more difficult for an individual with low English language proficiency relative to peers than tests with little to no verbal directions and nonverbal response options. For example, tests assessing verbal abilities, such as Comprehension on the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV; Wechsler, 2003), are more linguistically demanding than nonverbal tasks, such as Block Design on the WISC-IV, that assess fluid reasoning skills. Consequently, tests that are lower in linguistic demand may also provide a more accurate assessment of ELLs' cognitive abilities as a lower level of proficiency in English should not hinder performance on these measures (Flanagan et al., 2007).

The linguistic demand of cognitive measures has received more attention from U.S. lawmakers, researchers, and educators than the cultural content of such measures, but the attention has primarily focused on problems related to basic communication when working with students who have low proficiency levels with the English language. Results of a cognitive assessment would be invalid if the student is lacking in basic communication skills in English; however, a student's proficiency in basic interpersonal communication skills (BICS; Cummins,

1984) is not identical to the level of language proficiency necessary for a student to effectively communicate on cognitive or academic tasks (i.e., cognitive academic language proficiency [CALP]; Cummins, 1984). ELLs may display enough knowledge of the English language to adequately converse with teachers and peers, but they may not be able to communicate at the level expected of them in order to be successful academically. The misconception that ELLs should be able to function similarly in conversation as they do on school work can result in inappropriate assumptions that the students' academic difficulties are the result of a potential disability instead of language proficiency (Cummins, 1984). To undergo a special education evaluation due to a suspected disability, ELLs must first take a language proficiency test. However, such measures may not adequately identify weaknesses in English language proficiency that could negatively influence the results of their cognitive evaluation. Research has shown that increases in language proficiency correspond to increases in performance on a cognitive measure (Sotelo-Dynega, 2008). Thus, the potential for language proficiency to account for standardized cognitive assessment results cannot be overstated. It must be strongly considered during interpretation of findings and corresponding decisions regarding the presence of a disability. Limitations of available assessments, including use of measures that are considered culturally loaded and linguistically demanding, highlight the need for more appropriate methods for cognitive assessment of ELLs.

Purpose of the Study

Due to the increasing diversity in U.S. schools, cognitive assessment of English language learner students has increasingly become an area of significant concern for school psychologists. Inadequate (or lack of) training in working with ELLs, lack of proficiency in a second language, and limitations of currently available cognitive assessments are just a few of the challenges

school psychologists face in conducting appropriate psychological evaluations for ELLs. To address some of the difficulties in assessing ELL students, Flanagan et al. (2007) created (a) a cognitive test classification system and (b) a decision-making matrix.

The Culture-Language Test Classifications (C-LTC; Flanagan et al., 2007) and Culture-Language Interpretive Matrix (C-LIM; Flanagan et al.) were created to assist practitioners in accounting for the cultural and linguistic influences of commonly used cognitive measures developed and normed on samples of U.S.-born, native English speakers. The C-LTC is a classification system of commonly used cognitive measures based on levels of cultural and linguistic demand of individual subscales. The test classifications are then used in the C-LIM, a 3 (linguistic demand) x 3 (cultural loading) matrix, in which test scores can be entered and, cultural and linguistic influences can be assessed. Flanagan et al. (2007, 2013) contend that a diagonally declining pattern of test scores will be evident for ELLs without disabilities as a function of increases in cultural loading and linguistic demand. In essence, the presence of a diagonally declining pattern in test scores indicates that the student's cultural and language backgrounds are exerting considerable influences on the test scores; therefore, the scores should not undergo further interpretation (Flanagan et al.). Despite equivocal research in support of the validity of the C-LIM, school districts have adopted its use during evaluations of ELLs for special education services. Given the high-stakes nature of special education testing, it is imperative that the methods used be research-based and empirically supported.

The purpose of the present study was to test the utility of the C-LIM in distinguishing between test scores for Spanish-speaking, ELL students and monolingual, English-speaking students on the Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update (WJ- III-NU; Woodcock, McGrew, Schrank & Mather, 2007). The application of

diagnostic utility statistics to assess C-LIM decision-making is an innovative way to test the appropriateness of the C-LIM for use with the WJ-III-NU in practice. Only one previous study (Styck, 2012) has examined the diagnostic utility of the C-LIM for use on an individual basis with the WISC –IV; the findings indicated that the C-LIM's diagnostic accuracy was low. The aim of the present study was to examine the diagnostic accuracy of the C-LIM with another commonly used cognitive test battery, the WJ-III-NU. It is hoped that the findings of the present study will provide further evidence to consider when evaluating the utility of the C-LIM as a tool to inform decisions made about the validity of ELLs cognitive test scores and the use of such scores to inform high-stakes educational decisions.

LITERATURE REVIEW

The literature review is separated into seven main sections. The first section describes the Cattell-Horn-Carroll (CHC; McGrew, 2005) theory, followed by a section on its use within the framework of cross-battery assessment (XBA; Flanagan et al., 2007). The third section is devoted to the Culture-Language Test Classifications (C-LTC) and the fourth section is focused on the Culture-Language Interpretive Matrix (C-LIM; Flanagan et al., 2007). Both sections include the development of these extensions of XBA as well as their use during cognitive assessment of ELLs. The fifth and sixth sections provide a review of research conducted on the C-LTC and C-LIM, respectively, as well as limitations of the research. Section seven provides the rationale and purpose of this study, and resulting research question.

The Cattell-Horn-Carroll Theory

The Cattell-Horn-Carroll (CHC; McGrew, 2005) theory is a comprehensive taxonomy of cognitive abilities based upon a combination of Carroll's (1993) three-stratum theory, and Horn and Cattell's (1966) fluid-crystallized (Gf-Gc) theory of cognitive abilities. Carroll's three-stratum theory refers to (a) narrow, stratum I abilities, (b) broad, stratum II abilities, and (c) an overarching general stratum III ability, typically referred to as general intelligence or *g*, which were identified during a synthesis of over 460 factor-analytic studies on intelligence assessments in 1993. The stratum I abilities are considered to be facets corresponding to stratum II abilities as well as overall intelligence. For example, inductive and general sequential reasoning are stratum I abilities that fall within the broader ability of fluid intelligence, and fluid intelligence is subsumed within general intelligence.

Horn and Cattell's (1966) Gf-Gc theory was also developed through factor analysis and is based on the premise that fluid and crystallized intelligence are overarching general cognitive abilities under which multiple specific abilities are subsumed. Gf-Gc theory contains many of

the same broad and narrow abilities as delineated in Carroll's (1993) theory. The main difference between the two theories is the existence of general intelligence, a higher order factor that Horn and Cattell did not believe to be ultimately useful in practice and thus, not included in their theory.

Since the development of CHC theory, research has been conducted to further validate and refine the initial classifications of abilities as well as determine the fit of abilities described in CHC theory with tests contained in current cognitive batteries (Alfonso, Flanagan, & Radwan, 2005; Flanagan, 2000; Kaufman, Johnson, & Liu, 2008; Keith, Kranzler, & Flanagan, 2001; Phelps, McGrew, Knopik, & Ford, 2005; Taub & McGrew, 2004). The intelligence batteries associated with CHC include the Differential Ability Scales, Second Edition (DAS-II; Elliot, 2007), Kaufman Assessment Battery for Children, Second Edition (KABC-II; Kaufman & Kaufman, 2004), WJ-III-NU, and Stanford-Binet Intelligence Scales, Fifth Edition (SB5; Roid, 2003). However, most of these batteries do not include subscales that adequately represent of all of the broad ability areas delineated in the CHC theory (Flanagan et al., 2007). Consequently, if a comprehensive assessment guided by CHC theory is intended, the concurrent use of more than one intelligence battery is necessary to obtain measures of all broad ability areas that are considered reflective of the student's academic performance.

Cross-Battery Assessment

The practice of simultaneously using tests from multiple intelligence batteries is referred to as cross-battery assessment (XBA; Flanagan et al., 2007). Supporters (Alfonso et al., 2005; Floyd, Bergeron, McCormack, Anderson, & Hargrove-Owens, 2005) of CHC theory advocate use of this method to assess the range of broad and narrow abilities that could be influencing a student's academic performance in situations where the use of one cognitive measure does not tap all of the necessary areas. Research has supported the relations between seven broad

cognitive abilities contained within CHC theory and achievement in reading, math, and writing (Flanagan, Ortiz, Alfonso, & Mascolo, 2006; McGrew & Wendling, 2010). Flanagan and colleagues (2007) suggest that practitioners refer to such research when using XBA to ensure the best selection of measures that tap the broad abilities associated with the student's area of academic difficulty. Guidelines for conducting an XBA are as follows: (a) use the fewest number of cognitive batteries possible to obtain the necessary information (i.e., if one battery contains all necessary tests, only one battery should be used), (b) use tests classified by research conducted on CHC theory, whenever possible, and (c) use tests normed within a few years of each other (Flanagan et al., 2007).

However, there are potential problems with the practice of XBA. Glutting, Watkins, and Youngstrom (2003) raised nine concerns related to the use of XBAs:

> (1) the comparability of subtest scores obtained from different instruments, (2) the effects associated with modifying the presentation order of subtests, (3) issues of sampling and norming, (4) procedures used to group subtests into factors, (5) the use of ipsative score interpretation, (6) the extent of established external validity, (7) the relative efficiency and economy of the assessment process, (8) the vulnerability to misuse, and (9) the method in determining the correct number of factors to retain. (p. 361)

Similar concerns were also noted earlier by Watkins, Youngstrom, and Glutting (2002). Each of these issues is briefly described. Glutting and colleagues contend that the steady increase in participants' scores on the same cognitive measure over time, known as the Flynn effect (Flynn, 1999), could influence results and subsequently bias conclusions of studies in which scores from old and new versions of the same cognitive scale are combined. Also, administration of subtests

in a different order from which they were normed can produce order effects that have a varied influence on test scores depending on the student and the test situation (Glutting et al., 2003). Furthermore, no interrater reliability statistics were calculated during the initial classification of subtests based on the narrow ability definitions in CHC theory; thus, the consistency of these classifications are unknown (McGrew, 1997). Glutting and colleagues asserted that the few studies conducted on XBA used inadequate sampling and norming procedures, resulting in inadequate evidence necessary to establish the external validity of cross battery assessment. The lack of clear evidence in support of the reliability and validity of XBA was also cited as a reason for its misuse (Glutting et al., 2003). Glutting and colleagues also contested the use of ipsative score interpretation in XBA as research (Watkins, Kush, & Glutting, 1997) has not supported such practice. Finally, the last two concerns about cross-battery assessment, (i.e., the determination of the number of factors to retain and the efficiency of cognitive testing) are based on the lack of clear evidence supporting the unique benefits of XBA (Glutting et al., 2003). Given the high-stakes involved in psychoeducational testing and in the absence of unequivocal empirical evidence for using XBA, other assessment practices with more empirical support should be used instead.

Ortiz and Flanagan (2002a, 2002b) responded to each of Watkins and colleagues' (2002) concerns. One, in regard to comparability of test scores, Ortiz and Flanagan acknowledged the concern regarding the Flynn effect had merit, but only if practitioners used tests with norms more than 10 years old. They asserted that the norms should be comparable between tests with nationally representative normative samples as long as the tests were normed within 10 years of each other. Furthermore, Floyd and colleagues (2005) have provided support for the comparability of test scores across instruments. They found that the variance accounted for by

differences in test batteries was small (0% to 5%). Floyd and colleagues concluded that differences in scores across cognitive instruments that were not due to differences in ability were likely due to a combination of examinee characteristics and random error. Thus, the same factors (i.e., examinee characteristics such as fatigue and motivation as well as random error) must be considered as potential explanations for test scores for XBA evaluations as in traditional evaluations containing a single test battery.

Two, Ortiz and Flanagan (2002a) believed it was acceptable to use various subtests from different tests out of the standardized sequence, stating, "order effects are of little practical concern" (p. 32). However, the authors admitted that the administration of subtests out of sequence is "an under-researched topic" (Ortiz & Flanagan, 2002a, p. 32). Three, Ortiz and Flanagan also acknowledged that Watkins and colleagues' concern about categorizing many of the subscales according to CHC theory via logical rather than empirical means had merit. However, Ortiz and Flanagan countered this concern by citing a study completed after the initial CHC classification of subscales that supported the process. Caltabiano (2002; as cited in Flanagan et al., 2002) demonstrated a high rate of interrater agreement (approximately 96% for broad abilities) on the classifications. Four, Ortiz and Flanagan noted that ipsative score interpretation is no longer included as part of cross battery interpretation, and thus, is no longer a concern. Five, in regard to efficiency and economy of assessment, Ortiz and Flanagan indicated that Watkins and colleagues' assumptions that 18 subscales must be administered were incorrect (14 subscales is recommended) and that such a contention about XBA would not be defensible if legal action was taken about such an evaluation. Six, in regard to sampling and norming issues, Ortiz and Flanagan disagreed with the idea that large, nationally representative samples were necessary to validate XBA for use in practice. They stated, "constraining all clinicians to use a

single test to address any and all referral concerns is impractical and misguided" (Ortiz & Flanagan, p. 33). Finally, for the remaining concerns (i.e., extent of established external validity and vulnerability to misuse), Ortiz and Flanagan cited opinions from other psychologists (Horn, 1999; Kaufman, 2000; Prifitera, Weiss, & Saklofske, 1998) to support the validity of the XBA approach.

Despite the concerns raised, XBA has been generalized to the assessment of ELLs (Flanagan & Ortiz, 2007) through the Culture-Language Test Classifications (C-LTC; Flanagan et al.) and the Culture-Language Interpretive Matrix (C-LIM; Flanagan et al.). C-LTC was developed to sort tests based on level of cultural loading and degree of linguistic demand. C-LIM was created so that (a) classifications were more easily accessible for each test battery, and (b) cultural and linguistic influences inherent in tests could be methodically considered in determining the validity of cognitive test results for ELLs.

The Culture-Language Test Classifications

The Culture-Language Test Classifications (C-LTC; Flanagan & Ortiz, 2001) were developed to provide a practical method by which psychologists could determine the level of cultural influence or "loading" and level of language demand associated with frequently used cognitive measures when used for ELLs. The C-LTC is a table containing all classified tests from standardized cognitive batteries (e.g., WJ-III-NU) that have been categorized based on three levels of cultural loading and three levels of linguistic demand. The three levels for each variable are low, medium, and high. The following factors were considered during the classification process of the cultural loading of each test: (a) the content of the test (i.e., novel stimuli vs. culture-specific stimuli), (b) the focus of test (process used to solve a problem or the overall product), and (c) the interactions between the examiner and examinee (i.e., gestures apart from language). Similarly, classification of the level of language demand involved consideration

of the receptive and expressive language requirements for both the examiner and examinee (Flanagan et al., 2007).

When initially developed, the purpose of the C-LTC was to allow practitioners to select less culturally and linguistically biased tests from classified cognitive batteries that measured the broad ability areas of interest. Using tests with lower cultural and linguistic loadings is intended to result in less biased assessment of ELLs' cognitive functioning across the broad ability areas (Flanagan et al., 2007). However, several concerns with the C-LTC arose soon after its development. First, few ability areas could be adequately assessed if only tests with low language demand and cultural loading were selected. Second, the classification table became cumbersome when all possible test options in each category (e.g., low cultural loading and low linguistic demand, low cultural loading and medium linguistic demand, low cultural loading and high linguistic demand, etc.) were presented on the same page. Thus, to increase the ease in using the C-LTC, the amount of information contained in the test classification table was decreased. The information was split into multiple, test-specific classification tables so that examiners could choose a single battery (e.g., WISC-IV) and examine the cultural and linguistic loadings for only the tests contained therein. Third, finding measures of broad ability areas of interest that are also less culturally and linguistically biased is more difficult, if not in some cases impossible, to address as certain abilities inherently require acquisition of cultural and linguistic knowledge (e.g., crystallized intelligence [Gc]; Flanagan et al., 2007). Thus, to assess all potentially relevant cognitive ability areas that relate to a suspected disability and still address the potential influences of culture and language on a student's cognitive test performance, the Culture-Language Interpretive Matrix was developed.

Culture-Language Interpretive Matrix

The Culture-Language Interpretive Matrix (C-LIM; Flanagan & Ortiz, 2001; Flanagan et al., 2007, 2013) was developed to provide a way for psychologists to account for cultural and linguistic influences on cognitive scores obtained by ELLs. Although the C-LIM is often referred to in the singular form (i.e., matrix), it actually represents a set of matrices, each matrix depicting the subtest placement for a major test battery (e.g., WISC-IV or WJ-III-NU) based on C-LTC classifications. Figure 1 depicts where WJ-III-NU tests are placed within the C-LIM template. This template depicts how cultural loading increases from top to bottom (low to medium to high), and linguistic demand increases from left to right (low to medium to high). The C-LIM cultural loading levels differ based on vertical placement of tests in the matrix, while levels of linguistic demand differ based upon horizontal placement of tests in the matrix.

To use the C-LIM, standardized test scores from the administered cognitive battery are entered into the C-LIM template as categorized in the C-LTC. (A software program conceived by Flanagan et al. [2007, 2013] can be used for efficiency of C-LIM use and interpretation). If more than one test score is contained in the same cell, then an average standard score is computed for each cell. Flanagan and colleagues (2007) assert that a specific pattern will emerge in the C-LIM when students' cultural and linguistic backgrounds have affected their performance on a cognitive measure. The C-LIM pattern, in Figure 2, depicts how a typical ELL's performance on any cognitive measure is expected to be influenced by cultural and linguistic background in comparison to the norm group on which the standardized cognitive measure was developed (Flanagan & Ortiz, 2001, Flanagan et al., 2007; Flanagan et al., 2013). An ELL's test scores are expected to decrease (a) as the cultural loading of test scores increases (i.e., from top to bottom), (b) as the linguistic demand of the test scores increases (i.e., from left to right), and (c) as combined influences of cultural loading and language demand on the test scores increase

Degree of Linguistic Demand

Degree of Cultural Loading	Low		Medium		High	
	Test Name	Score	Test Name	Score	Test Name	Score
Low	Spatial Relations	___	Numbers Reversed	___	Analysis-Synthesis	___
			Visual Matching	___	Concept Formation	___
	Cell Average	___	**Cell Average**	___	**Cell Average**	___
Medium	Picture Recognition	___	Retrieval Fluency	___	Auditory Attention	___
			Visual Auditory Learning	___	Decision Speed	___
					Memory for Words	___
					Sound Blending	___
	Cell Average	___	**Cell Average**	___	**Cell Average**	___
High					General Information	___
					Verbal Comprehension	___
	Cell Average	___	**Cell Average**	___	**Cell Average**	___

Figure 1. C-LIM with WJ-III-NU test names. Adapted from *Essentials of Cross-Battery Assessment, 2nd Edition* (p. 189), by D. P. Flanagan, S. O. Ortiz, and V. C. Alfonso, 2007, New Jersey: Wiley & Sons, Inc. Copyright 2007 by Wiley & Sons, Inc. Adapted with permission.

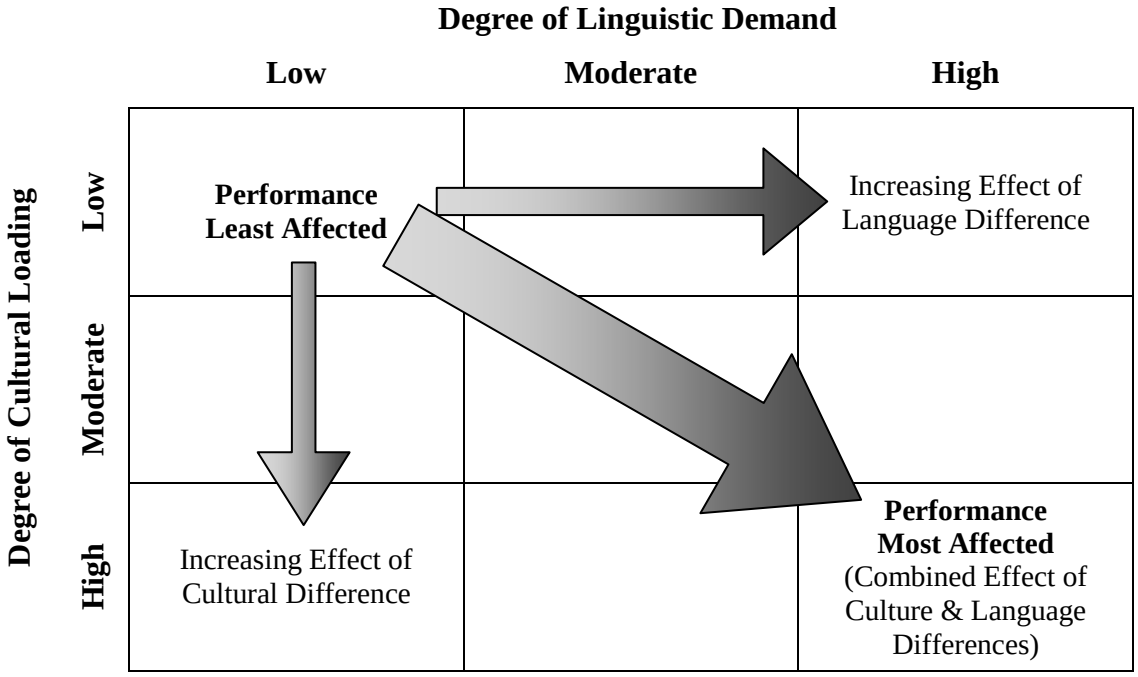

Figure 2. C-LIM demonstrating the pattern of expected performance for culturally and linguistically diverse students. From *Essentials of Cross-Battery Assessment, 2nd Edition* (p. 177), by D. P. Flanagan, S. O. Ortiz, and V. C. Alfonso, 2007, New Jersey: Wiley & Sons, Inc. Copyright 2007 by Wiley & Sons, Inc. Reprinted with permission.

(i.e., from top left to bottom right of the C-LIM).

If the software provided by Flanagan and colleagues (2013) is used for interpretation of the C-LIM pattern, a set of three bar graphs accompany the matrix. Each bar represents a cell in the C-LIM (9 total), and the bars are ordered from left to right based on increases in cultural loading and linguistic demand. The first bar graph represents the combined influence of cultural loading and linguistic demand (from low to moderate to high). The remaining two graphs represent the influences of (a) cultural demand only (low to moderate to high), and (b) linguistic demand only (low to moderate to high). As an example, a graph of scores that follow a declining pattern along the combined increase in cultural loading and linguistic demand is presented in Figure 3.

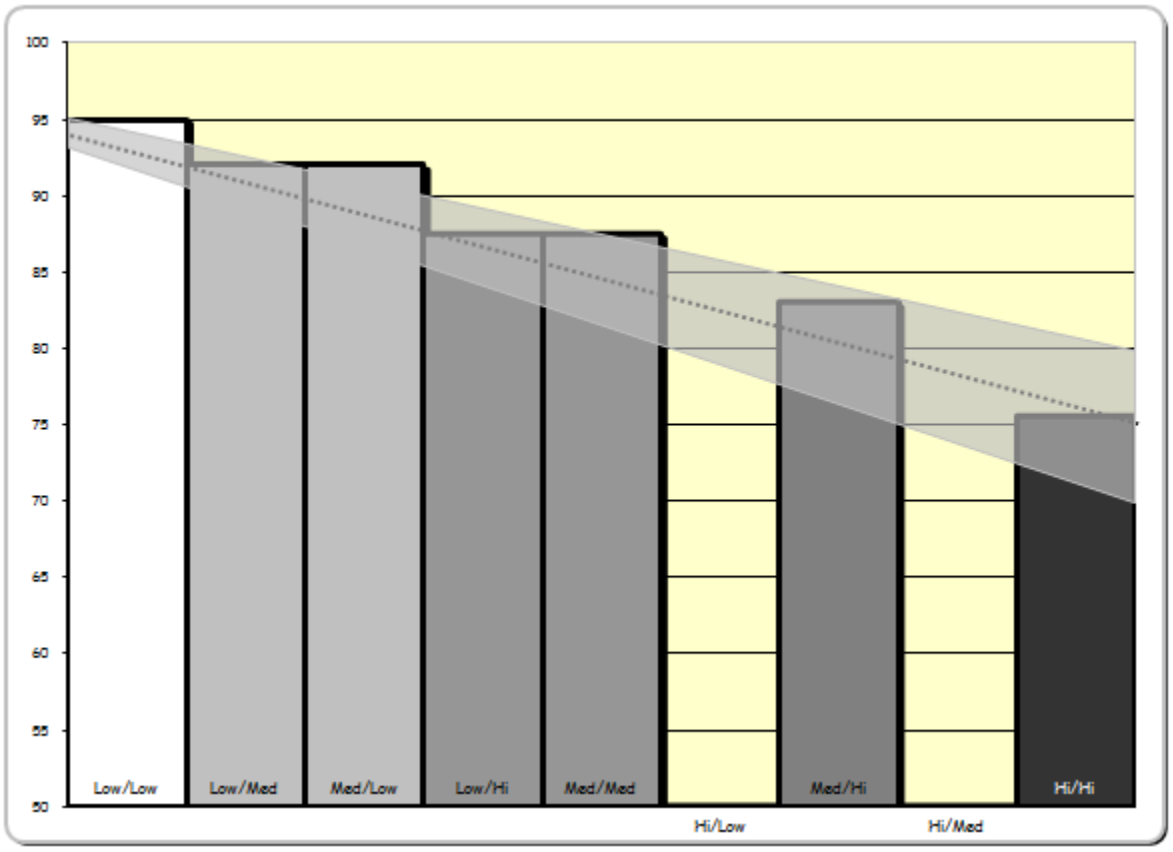

Figure 3. Bar graph representing decline in cognitive test scores based on cultural loading and linguistic demand. Adapted from *Essentials of Cross-Battery Assessment, 2nd Edition* (p. 195), by D. P. Flanagan, S. O. Ortiz, and V. C. Alfonso, 2007, New Jersey: Wiley & Sons, Inc. Copyright 2007 by Wiley & Sons, Inc. Adapted with permission.

Flanagan and colleagues (2007) indicate that differing degrees of language proficiency and acculturation will result in differing amounts of deviation from the norm, which is defined as an average individual taking the test (i.e., Standard Score = 100). Therefore, Flanagan and colleagues outlined three different scoring patterns to assist practitioners in determining the degree of attenuation in scores that could be expected for an individual whose cultural and linguistic background differs from a native English speaker who was born in the U.S. The first scoring pattern of *Slightly Different* is used to describe an individual who exhibits a high level of English language proficiency and a high level of acculturation. For example, a person who has lived in the U.S. for at least seven years and is fluent in conversational English but demonstrates some difficulties with academic English would be categorized as *Slightly Different*. A person who would be described as *Moderately Different* demonstrates an intermediate level of English language proficiency and a moderate level of acculturation. Such person has lived in the U.S. for three to seven years and demonstrates some difficulties with conversational English and limited proficiency in academic English. Finally, an individual who would be described as *Markedly Different* has limited proficiency with the English language and still has much to learn about U.S. culture. A person in this category struggles with conversational English and has typically lived in the U.S. for a short period of time. Figure 4 shows the expected decline in scores for Slightly, Moderately, and Markedly Different categories.

As shown in Figure 4, for a single category (i.e., *Slightly Different*, *Moderately Different* or *Markedly Different*), there are only five possible score ranges represented in the matrix. When cells with the same score range are grouped together, it results in the five-level model shown in Figure 5. Thus, for example, when a student categorized as *Slightly Different* completes subscales classified as having low cultural loading/medium linguistic demand and

Degree of Cultural Loading	Degree of Linguistic Demand: Low	Medium	High
Low	Slightly Different: 3-5 points Moderately Different: 5-7 points Markedly Different: 7-10 points	Slightly Different: 5-7 points Moderately Different: 7-10 points Markedly Different: 10-15 points	Slightly Different: 7-10 points Moderately Different: 10-15 points Markedly Different: 15-20 points
Medium	Slightly Different: 5-7 points Moderately Different: 7-10 points Markedly Different: 10-15 points	Slightly Different: 7-10 points Moderately Different: 10-15 points Markedly Different: 15-20 points	Slightly Different: 10-15 points Moderately Different: 15-20 points Markedly Different: 20-25 points
High	Slightly Different: 7-10 points Moderately Different: 10-15 points Markedly Different: 15-20 points	Slightly Different: 10-15 points Moderately Different: 15-20 points Markedly Different: 20-25 points	Slightly Different: 15-20 points Moderately Different: 20-30 points Markedly Different: 25-35 points

Figure 4. General guidelines for *Slightly*, *Moderately*, and *Markedly Different* culturally and linguistically diverse individuals. From *Essentials of Cross-Battery Assessment, 2nd Edition* (p. 200), by D. P. Flanagan, S. O. Ortiz, and V. C. Alfonso, 2007, New Jersey: Wiley & Sons, Inc. Copyright 2007 by Wiley & Sons, Inc. Reprinted with permission.

Degree of Cultural Loading	Degree of Linguistic Demand: Low	Medium	High
Low	Level 1	Level 2	Level 3
Medium	Level 2	Level 3	Level 4
High	Level 3	Level 4	Level 5

Figure 5. C-LIM collapsed into five levels based upon increases in cultural loading and linguistic demand.

medium cultural loading/low linguistic demand, the expectation (based on Figure 3) is that his or

her scores on subscales under each classification would be five to seven points lower than what is considered the normal range. Because the expected difference in scores is the same for both classifications (low cultural loading/medium linguistic demand and medium cultural loading/low linguistic demand), these cells are collapsed into a single level within the 5-level model.

Summary and Critique of C-LTC & C-LIM

The C-LIM was designed to provide a methodical approach toward investigation of cultural and linguistic influences on the cognitive scores of culturally and linguistically diverse students. However, there are limitations in its delineation, potential use, and interpretation. One, many of the C-LTC and subsequent C-LIM classifications were determined through an expert consensus procedure, but no explanation of the reasoning behind the classifications for each test or the actual consensus procedure was given (Flanagan et al., 2007). Other than indicating that expert consensus was used, no details regarding the qualifications of the experts who were consulted, the aspects of the tests that were considered, the level of agreement (e.g., 100 percent versus majority) necessary to classify a test in a certain category (e.g., high cultural loading, low linguistic demand) or any other information related to the process have been offered in the publications available on the C-LTC and C-LIM. This limitation is consistent with concerns expressed by Watkins and colleagues (2002) about the development of XBA (e.g., the lack of interrater agreement).

Two, the few classifications that were based on previous research were restricted to a small number of outdated studies (Cummins, 1984; Goddard, 1917; Jensen, 1974, 1976; Mercer, 1979; Sánchez, 1934; Valdés & Figueroa, 1994). Thus, further investigation is needed to validate the initial findings. The findings are also based on outdated measures, such as the WISC-R (Cummins, 1984; Valdés & Figueroa, 1994) and were completed on samples (Goddard,

1917; Mercer, 1979; Sanchez, 1934) unrepresentative of the level of cultural and linguistic diversity evident in U.S. schools today.

Finally, Flanagan and colleagues (2007) offer general guidelines for expected patterns of performance (i.e., ranges of expected standard scores) identified as *Slightly, Moderately, and Markedly Different* (based on the cultural and linguistic background of the students), but do not explain how the ranges were established. It is unclear whether the ranges were based directly on research, were extrapolated from scores obtained in research, or were developed in another manner.

Research on the C-LTC and C-LIM

Research conducted on the C-LTC and C-LIM is sparse. To date, eleven studies have been conducted: ten unpublished doctoral dissertations (Aziz, 2009; Brown, 2008; Cormier, 2012; Dhaniram-Beharry, 2008; Lella Souravlis, 2010; Nieves-Brull, 2006; Styck, 2012; Templeton, 2012; Tychanska, 2009; Verderosa, 2007) and one published study (Kranzler, Flores, & Coady, 2010). Six of the dissertations were completed at the same university. Brief summaries of the research questions, samples, analyses, findings and conclusions of the eleven studies are provided below, followed by a summary table presenting details for each study, and a discussion of common limitations across studies.

Research Questions

The content of research questions addressed in prior studies primarily focused on whether the declining C-LIM pattern was evident in a sample of ELLs (Aziz, 2009; Brown, 2008; Kranzler et al., 2010; Lella Souravlis, 2010, Nieves-Brull, 2006; Styck, 2012; Tychanska, 2009; Verderosa, 2007). In addition, monolingual, English-speaking students (MESs) were used as a comparison group in some studies (Aziz, 2009; Lella Souravlis, 2010, Nieves-Brull, 2006; Styck, 2012; Tychanska, 2009). Aziz (2009) and Nieves-Brull (2006) focused on a comparison of

ELLs' obtained subscale scores on a given cognitive test to the expected scores specified in the C-LTC classifications in their respective studies. However, more studies (Kranzler et al., 2010; Lella Souravlis, 2010; Styck, 2012; Tychanska, 2009; Verderosa, 2007) examined a general pattern of decline in subscale scores (without testing specific score ranges) for ELLs. Also, investigated was the pattern of decline for ELLs with global cognitive impairment (GCI; Aziz, 2009; Lella Souravlis, 2010), ELLs with speech or language impairment (SLI; Lella Souravlis, 2010; Tychanska, 2009), or ELLs with specific learning disability (SLD; Styck, 2012; Tychanska, 2009).

A couple of researchers (Cormier, 2012; Verderosa, 2007) examined the influences of culture and language on test scores and Dhaniram-Beharry (2008) investigated the C-LIM pattern as a function of race, gender, grade level, and disability status. Some studies (Brown, 2008; Cormier, 2012; Nieves-Brull, 2006; Templeton, 2012) focused on identifying appropriate placement of subscales in the C-LIM for the test battery administered. Brown (2008) and Templeton (2012) used tests in their respective studies that have not been specifically classified in the C-LIM. Specifically, Brown used the Batería III in the C-LIM, as it is the Spanish counterpart of the WJ-III-NU. Templeton, on the other hand, used the performance of ELLs on the Delis-Kaplan Executive Function System (D-KEFS) to identify potential C-LTC classifications for the test.

The purpose for most of the prior research (Aziz, 2009; Brown, 2008; Cormier, 2012; Dhaniram-Beharry, 2008; Kranzler et al., 2010; Lella Souravlis, 2010; Nieves-Brull, 2006; Templeton, 2012; Tychanska, 2009; Verderosa, 2007) was to examine the C-LTC or C-LIM as it functioned on a group level. Only Styck (2012) focused on whether the C-LIM adequately differentiated between individuals who should or should not follow a declining pattern.

Samples

In previous research on the C-LIM, most studies involved ELLs (Aziz, 2009; Brown, 2008; Kranzler et al., 2010; Lella Souravlis, 2010; Nieves-Brull, 2006; Templeton, 2012; Tychanska, 2009; Verderosa, 2007; Styck, 2012), and some studies used MESs as participants (Aziz, 2009; Cormier, 2012; Dhaniram-Beharry, 2008; Lella Souravlis, 2010; Nieves-Brull, 2006; Tychanska, 2009; Styck, 2012). A few studies used non-referred samples (Brown, 2008; Cormier, 2012; Kranzler et al., 2010, Templeton, 2012), but most samples consisted of students referred for special education evaluations (Aziz, 2009; Dhaniram-Beharry, 2008; Lella Souravlis, 2010; Nieves-Brull, 2006; Tychanska, 2009; Verderosa, 2007; Styck, 2012). In three studies, the sample consisted entirely of students identified with disabilities (Aziz, 2009; Lella Souravlis, 2010; Tychanska, 2009).

Many prior studies used samples containing fewer than 200 cases (Brown, 2008; Dhaniram-Beharry, 2008; Kranzler et al., 2010; Templeton, 2012; Verderosa, 2007). Furthermore, the largest number of ELLs speaking the same language contained in a single sample was 86. In general participants ranged in age from 2 to 56; however, most studies (Brown, 2008; Cormier, 2012; Dhaniram-Beharry, 2008; Kranzler et al., 2010; Nieves-Brull, 2006; Templeton, 2012; Tychanska, 2009; Styck, 2012) used samples that consisted entirely of school-aged individuals (ranging in age from 5 to 18 years). In regard to gender, there were generally more males (50% -68%) than females (32%-50%) represented in the prior studies. Also, race and ethnicity data were not reported in 7 of the 11 prior investigations on the C-LIM (Aziz, 2009; Brown, 2008; Cormier, 2012; Kranzler et al., 2010; Styck, 2012; Templeton, 2012; Verderosa, 2007). For studies in which race/ethnicity information was reported, Hispanic and White individuals were represented in two studies (Nieves-Brull, 2006; Tychanska, 2009), one

used a sample consisting of Asian, Hispanic, and White individuals (Lella Souravlis, 2010), and one sample consisted of African American, Hispanic, and White individuals (Dhaniram-Beharry (2008).

Statistical Analyses

Most of the research on the C-LIM used analyses of group means via MANOVA (Nieves-Brull, 2006), ANOVA (Kranzler et al., 2010; Nieves-Brull, 2006; Verderosa, 2007; Styck, 2012), or t tests (Aziz, 2009; Dhaniram-Beharry, 2008; Kranzler et al., 2010; Lella Souravlis, 2010; Nieves-Brull, 2006; Styck, 2012; Tychanska, 2009; Verderosa, 2007). A few researchers used slight variations of analyses to examine group means, such as Euclidean distance calculations (Dhaniram-Beharry, 2008), or analysis of mean squared differences (Nieves-Brull, 2006). In addition, several researchers either did not use inferential statistics to study the C-LIM pattern (Brown, 2008; Templeton, 2010), or supplemented inferential statistics by doing a visual inspection of group means and calculating frequency counts (Kranzler et al., 2010; Nieves-Brull, 2006). One researcher, Cormier (2012), used latent variable structural equation modeling to study the influences of culture and language ability on cognitive test performance based on the classification in the C-LIM. Another researcher, Styck (2012), used diagnostic utility statistics to examine the diagnostic accuracy of the C-LIM for ELLs, focusing on the impact at an individual level instead of at a group level.

Findings and Conclusions

Findings of previous research on the C-LTC and C-LIM have been mixed. On a group level, no study has provided consistent support for the C-LTC classifications or C-LIM interpretation. Nieves-Brull (2006) and Verderosa (2007) provided some support, although not consistent for the pattern of decline in the C-LIM. Studies conducted by Aziz (2009), Brown

(2008), Kranzler and colleagues (2010), and Styck (2012) had findings that were not in support of C-LIM interpretation. However, despite inconsistent findings, several researchers concluded that their results supported the use of the C-LIM (Aziz, 2009; Lella Souravlis, 2010, Tychanska, 2009). In addition, due to differences in results from the expected C-LIM pattern identified by the C-LTC classifications, several researchers recommended re-classification of subscales in the C-LTC (Brown, 2008; Cormier, 2012; Nieves-Brull, 2006). Tychanska suggested that problems with the sample used (i.e., misdiagnosis of ELLs with SLD) could have produced fewer findings in support of the C-LIM than expected. Cormier indicated that fewer findings in support of the C-LIM in his study may have been due to weaknesses in the measures of culture or language used to investigate the C-LIM.

Results have been mixed for the influence of linguistic demand on ELLs' test scores (Cormier, 2012; Kranzler et al., 2010; Tychanska, 2009; Verderosa, 2007). Cormier (2012) found that the linguistic demand variable operated as expected given the pattern outlined in the C-LIM, but Tychanska (2009) and Verderosa (2007) reported mixed findings for the influence of linguistic demand on average ELL test scores. Kranzler and colleagues (2010) found that the effect of linguistic demand on ELLs average test scores was not statistically significant. For acculturation, results of prior studies were also mixed (Cormier, 2012; Kranzler et al., 2010; Verderosa, 2007). Cormier and Verderosa found that the impact of cultural loading on ELLs test performance was negligible, but Kranzler and colleagues reported a statistically significant effect of cultural loading on ELLs' test scores and mixed post hoc findings.

Dhaniram-Beharry (2008) found that the expected pattern of decline in scores for ELLs was not evident in an MES sample as a function of race, gender, grade level, or disability status. Dhaniram-Beharry concluded that this finding supported C-LIM interpretation because it is

intended to differentiate between ELLs based on disability status, not between English speakers based on race, gender, grade level, or disability status. Templeton (2012) reported that differences between ELLs and the norm sample were evident, as hypothesized. Templeton used the scores from ELLs to suggest potential C-LTC classifications for the Delis-Kaplan Executive Function System (D-KEFS).

In examining the C-LIM pattern at an individual level, Kranzler and colleagues (2010) reported that 30-43% of non-referred ELLs had test scores that followed one of the three possible patterns of decline (cultural loading only, linguistic demand only, or combined cultural loading and linguistic demand), while an almost equal percent (41%) of ELLs had test scores that did not follow any of the three patterns. Styck (2012) used diagnostic utility statistics to determine how well the presence of the diagonally declining (3-cell) pattern distinguished between MESs, who were not expected to follow the pattern, and ELLs who were (if non-SLD) or were not (if SLD) expected to follow the pattern. Styck's findings indicated that the C-LIM adequately detected MESs (from the WISC-IV normative sample), who did not have a declining pattern in performance (i.e., specificity = .90-.95, guideline $\geq$ .80), but poorly detected ELLs with a declining pattern (sensitivity = .00-.10, guideline $\geq$ .70). Furthermore, ROC curve analysis and AUC values indicated that the C-LIM demonstrated low diagnostic accuracy (i.e., between 0.5 and 0.7, according to Swets' [1988] guidelines) in distinguishing between (a) ELLs with SLD and the WISC-IV normative sample, (b) ELLs without SLD and the WISC-IV normative sample, and (c) ELLs with and without SLD. Thus, the C-LIM was not able to accurately differentiate between those who were supposed to follow a declining pattern and those who were not. Kranzler and colleagues, and Styck concluded that their findings did not support use of the C-LIM to inform decisions regarding the validity of ELLs cognitive test scores.

Prior research on the C-LTC and C-LIM provides "equivocal support for their use, at best" (Kranzler et al., 2010; p. 435). Research investigating the C-LTC and C-LIM on a group level has been mixed, with more studies producing non-significant results than expected. Few researchers have analyzed the effect of cultural and linguistic factors on ELLs' cognitive performance in the C-LIM, and those researchers who have examined these factors have had mixed results. Furthermore, even fewer researchers have investigated the use of the C-LIM on an individual level, as it is meant to be used. The findings of investigations on an individual level suggest that the expected C-LIM pattern for ELLs is not demonstrated the extent expected based on Flanagan et al.'s assertions (Flanagan et al., 2007, 2013). Table 1 provides a summary of each of the studies that have been conducted on the C-LTC or C-LIM. The sample characteristics, measure and models tested, analyses used, and results of each of the studies are presented.

Limitations of C-LTC and C-LIM Research

In addition to concerns about the development of the C-LTC and C-LIM, which were discussed earlier (pp. 24-25), limitations of the existing research further call into question the use of the C-LTC and C-LIM for the purposes for which they were developed.

Sampling Procedures

Many of the C-LIM studies (Aziz, 2009; Dhaniram-Beharry, 2008; Nieves-Brull, 2006; Tychanska, 2009; Verderosa, 2007, Styck, 2012) used a sample of students referred for evaluations to determine disability status and subsequent special education eligibility. Without knowing the disability status of the participants (i.e., Verderosa, 2007), no clear conclusion about the usefulness of the C-LIM can be made. To test the C-LIM requires delineating the status of the participants, by language, cultural background, and disability status, as the expected scores are based on this information. Small sample size is also a limitation of existing investigations on

Table 1

Summary of Research Conducted on the C-LTC and C-LIM in Chronological Order of Publication Date

Study	Sample	Measures/Models Tested	Analyses	Results
Nieves-Brull (2006)	N = 119; MES = 53, BESS = 66 Age: 6-16 years 55% Hispanic, 45% Caucasian Referred but ineligible for special education	Primary: WISC-III Models*: 1-Scores = 100, 2-Scores = 85, 3-Scores = value associated with C-LTC, 4-Scores = value associated with Alternative model	MANOVA, ANOVA, independent sample t tests, visual inspection paired sample t tests, effect sizes, frequency counts	Group level: Mean differences in cognitive performance between MES and BESS groups ($p < .001$, eta = .46), with MES scoring higher on 6 of 10 subtests ($p < .000$ to .01). MSDs were sig. different across MES and BESS groups for 85 ($p < .001$, $d = 0.78$), C-LTC ($p < .001$, $d = 0.91$), and Alt ($p < .001$, $d = 0.92$) models. Smallest effect sizes, indicating best model fit, noted for 85 ($d = 2.22$) and Alt ($d = 2.22$) models for MES, and 85 model ($d = 2.09$) for BESS. Individual level: Frequency at which scores reached the smallest MSD: 100 for MES ($n = 37$, 70% of group), Alt for BESS ($n = 23$, 35% of group).
Verderosa (2007)	N = 60, all BESS Age: 3-5 years Referred for evaluation for special education	Primary: DAS Secondary: Home Language Survey, Bidimensional Acculturation Scale	Paired sample t tests, repeated one-way ANOVAs, independent sample t tests	Group level: Mean differences noted in 13 out of 14 DAS subtest comparisons ($p = .000$), with all sig. differences in expected direction based on classifications in C-LIM. Students who spoke some English at home generally obtained higher DAS scores than students who spoke only Spanish ($p = .002$ to .278, post-hoc analyses were n.s.). Non-significant effect of acculturation on DAS scores ($p = .069$ to .714).
Brown (2008)	N = 35, all BESS Age: 8-10 years Non-referred, regular education students	Primary: Batería III Secondary: Language Assessment Scales, Acculturation Quick Screen	Visual inspection	Group level: Comparisons made using single cut score (pulled from ranges in Figure 3, p. 19) and 5-level system (Figure 4, p. 20) At levels 1, 2, and 4, obtained mean scores were higher than assigned C-LIM values. At levels 3 and 5, obtained mean scores were lower than assigned C-LIM values.

Note. MES = monolingual, English-speakers, BESS = bilingual, English/Spanish-speakers, WISC-III = Wechsler Intelligence Scales for Children – Third Edition, DAS = Differential Ability Scales. * Mean squared differences (MSDs) were calculated between obtained and predicted scores for each model.

Table 1 (continued)

Study	Sample	Measures/Models Tested	Analyses	Results
Dhaniram-Beharry (2008)	N = 64, all MES Grades: K – 8th 37% African American, 33 % Caucasian, 30% Hispanic All diagnosed with SLD	Primary: WISC-III, WISC-IV, & WJ-III	Independent samples t tests, Euclidean distance calculations	Group level: Investigated differences between obtained and identified means (based on 5-level system, p. 20) based on (a) race, (b) disability, (c) gender, (d) grade, and (e) cognitive battery used. Expected to find that mean scores did not follow the declining C-LIM pattern (i.e., n.s. results). Mean score differences found between C-LIM levels 4 and 5 across several grade, ethnicity and gender comparisons (p = .000 to .025). All other results were n.s.
Aziz (2009)	N = 142; MES = 74, bilingual = 68 (12 lang. other than English) Age: 3-56 years All with GCI *	Primary: SB-V, WAIS-III, & WISC-IV Model: Adjusted (to account for cognitive deficiency) 5-level system using single cut score per level: L1 = 73, L2 = 71, L3 = 68, L4 = 65, L5 = 60.	One-sample t tests	Group level: On average, full scale and C-LIM cell average scores were lower than average standard score (M = 100; p < .001). For MES group, obtained means were higher than C-LIM cut scores at levels 1, 3, 4, and 5 (p = .000 to .04). For bilingual group, obtained scores were lower than C-LIM cut scores at levels 1, 2, 3, and 4 (p = .000 to .02), and higher than the C-LIM cut score at level 5 (p = .001). Bilingual group assessed with WISC-IV scored equal to or higher than C-LIM cut score, while bilingual groups assessed with SB-V or WAIS-III scored lower than the C-LIM cut score. Diagonal decrease in effect size evident for both bilingual and MES groups.

Note. NLD = Batería III = Batería – Third Edition, WISC-IV = Wechsler Intelligence Scales for Children – Fourth Edition, WJ-III = Woodcock Johnson, Tests of Cognitive Abilities – Third Edition, SLD = Specific Learning Disability, GCI = Global Cognitive Impairment, SB-V = Stanford Binet – Fifth Edition, WAIS-III = Wechsler Adult Intelligence Scale – Third Edition.

* GCI was reflected in the following disabilities: autism, pervasive developmental disorder, not otherwise specified, mental retardation, Down syndrome, or Fragile X syndrome.

Table 1 (continued)

Study	Sample	Measures/Models Tested	Analyses	Results
Tychanska (2009)	N = 182; MES = 79, BESS = 86, BEIS = 15 Age: 5-15 years All diagnosed with SLD (n = 115) or SLI (n = 67)	Primary: WISC-III, WISC-IV, & WPPSI-III	Independent samples t tests, one sample t tests	Group level: MESs with SLD had higher subtest scores than MESs with SLI (p = .01), and ELLs with SLD scored lower than MESs with SLD (p = .001). Comparisons between average subtest scores of ELLs with SLD or SLI and ELLs or MESs with SLI were n.s.. For MESs with SLD, 3 of 7 comparisons of C-LIM cell averages to Standard Score of 100 were stat. sig. (p = .000 to .024, d = -0.17 to -0.78). For ELLs with SLD, contrary to expectations, all 7 comparisons of C-LIM averages to Standard Score of 100 were stat. sig. (p = .000 to .008, d = -0.43 to -1.04), and scores decreased along C-LIM diagonal (L1 = 93, L3 = 89, L5 = 84). For ELLs with SLI, stat. sig. differences in 4 of 7 cells with moderate to high linguistic demand (one-sample t test, p = .000), and effect sizes increased along increases in linguistic demand (d = -0.46 to -1.38). For MESs with SLI, stat. sig. differences between 4 obtained scores and 100 were stat. sig. (p = .000 to .013), with effect size increases along increases in linguistic demand (d = -0.42 to -0.76).
Lella Souravlis (2010)	*N* = 117; MES = 62, bilingual = 55 (20 lang. other than English) Age: 2-12 years 53% White, 41% Asian, & 6% Hispanic All diagnosed with GCI or SLI	Primary: WPPSI-III & WISC-IV Model: 5-level C-LIM (p. 20)	Paired sample t tests, independent samples t tests	Group level: Comparisons between obtained scores and C-LIM cut scores for ELLs with SLI (p = .000 to .637) and for MESs with SLI were mixed (p = .007 to .909). Comparisons of scores from ELLs with SLI and MESs with SLI were also mixed (p =.000 to .953, d = -0.01to 2.27). Comparisons between obtained scores and C-LIM cut scores for ELLs and MESs with CGI were mixed (p = .001 to .979). Comparisons of scores from ELLs with CGI and MESs with CGI were also mixed (p = .000 to .802, d = -1.57 to .54). Results do not provide clear support for a differential C-LIM pattern between MESs and ELLs with CGI.

Note. BESS = bilingual, English/Spanish speaker, BEIS = bilingual, English/Italian-speaking, SLI = Speech or Language Impairment, WPPSI-III = Wechsler Preschool and Primary Scale of Intelligence – Third Edition.

Table 1 (continued)

Study	Sample	Measures/Models Tested	Analyses	Analyses/Results
Kranzler, Flores, & Coady (2010)	*N* = 46, all bilingual (17 lang. other than English) Age: 5-18 years 22 countries represented (*n* = 1 born in U.S.) 43% free/reduced eligible Non-referred, regular education students	Primary: WJ-III Secondary: Comprehensive English Language Learning Assessment (Correlations between language proficiency and cognitive ability were n.s.) Model: 1-*combined effect* of cultural loading and linguistic demand, 2-high *linguistic demand* across low, moderate, and high levels of cultural loading, 3-moderate *cultural loading* across all levels of linguistic demand	Within-subjects ANOVAs, repeated ANOVAs, paired sample t test, visual inspection	Group level: The effects of increasing linguistic demand (p = .202) or increasing cultural loading (p = .199) on cognitive performance were n.s. The *combined effect* on cognitive performance was statistically and practically sig. (p = .039, partial eta^2 = 0.38). Post hoc comparisons were also sig. Effect of *linguistic demand* on subtest mean scores resulted in a statistically and practically sig. effect (p = .001, partial eta^2 = 0.32). Post hoc results were inconsistent. Effect of *cultural loading* on subtest mean scores was statistically and practically sig. (p = .001, partial eta^2 = 0.25), but post hoc results were mixed. Individual level: 30-43% of participants followed one of the three predicted C-LIM patterns (1, 2, or 3), and 13 % followed all three patterns. 41% of participants had scores that did not follow any of the predicted patterns.
Templeton (2012)	*N* = 30, all BESS Age: 8-13 years Non-referred, regular education students	Primary: D-KEFS Secondary: California English Language Development Test	Visual inspection, qualitative analysis	Group level: Examination of group means in order to classify D-KEFS subtests by cultural loading and linguistic demand. C-LIM classifications were determined during this study, not tested.

Note. WJ-III = Woodcock-Johnson Tests of Cognitive Abilities, Third Edition, D-KEFS = Delis-Kaplan Executive Function System.

Table 1 (continued)

Study	Sample	Measures/Models Tested	Analyses	Results
Cormier (2012)	*N* = 4,404 (subset of nationally representative, WJ-III-NU norm sample) Age: 7-18 years	Primary: WJ-III-NU	Latent variable structural equation modeling	Model (Phase) :1- examination of theoretical relationships between cultural loading and linguistic demand and cognitive test performance, 2- split linguistic demand into receptive and expressive language and investigated the relationship between receptive and expressive language skills and cognitive test performance. Group level: Results were similar across age groups for Phase 1 and Phase 2. Path coefficients between latent variables of cultural loading and individual test performance were n.s. Path coefficients between latent variable of linguistic demand and individual test performance on General Information (*r* = .79), Concept Formation (*r* = .82), and Verbal Comprehension (*r* = .90) were significant.
Styck (2012)	*N* = 2,119; ELL = 86, MES = 2,033 Age: 6-16 years All ELLs were referred for a special education evaluation. All MES were not referred.	Primary: WISC-IV Secondary: WIAT-III, WJ-III	One-way ANOVA, sensitivity, specificity, AUC, and ROC Analyses	Group level: All average WISC-IV scores on subtests, indices, and FSIQ were significantly different between ELLs and MESs (*p* = .001 to .002). Individual level: Low frequency of true positive (ELLs identified as different) and high frequency of true negative (MES identified as disordered) decisions in comparisons between ELLs and students from the WISC-IV norm sample.AUC values ranging from .51 to .53 for comparisons between ELLs with SLD and students in the WISC-IV norm sample.AUC values ranged from .46 to .53 irrespective of C-LIM cut score or SLD criteria used.

Note. WJ-III-NU = Woodcock Johnson, Tests of Cognitive Abilities – Third Edition, Normative Update, ELL = English language learner, MES = monolingual English speaker, WISC-IV = Wechsler Intelligence Scale for Children, Fourth Edition, WIAT-III = Wechsler Individual Achievement Tests, Third Edition.

the C-LTC or C-LIM. All studies except those that used normative data (i.e., Cormier, 2012; Styck, 2012) used samples of fewer than 200 cases and half of the studies on the C-LIM used samples of less than 100 (Brown, 2008; Dhaniram-Beharry, 2008; Kranzler et al., 2010; Verderosa, 2007; Templeton, 2012).

Another sampling concern was the lack of appropriate comparison groups. Some studies (Aziz, 2009; Dhaniram-Beharry, 2008; Lella Souravlis, 2010; Tychanska, 2009) had samples of ELLs and/or monolingual English speaking (MES) students, who had been identified as having some type of disability, but there were no control samples such as ELLs and MES without disabilities, for comparison. Using existing data from psychoeducational evaluations limits the possibility of having control groups, as was the case in these studies. At minimum, the sample could have included students who had been evaluated but not identified with any disabilities.

Three studies (Brown, 2008, Verderosa, 2007, Templeton, 2012) did not include a monolingual, English speaking (MES) sample as a comparison group. MES samples are necessary to establish the viability of C-LIM. To verify that the declining pattern observed for ELLs is valid, other alternatives must be ruled out; one is that no or virtually few non-ELLs should show the declining pattern of cognitive test scores as cultural loading and linguistic demand of the measures increase.

The geographical location where the studies were conducted is also a limitation. A majority of the studies (Aziz, 2009; Dhaniram-Beharry, 2008; Lella Souravlis, 2010; Nieves-Brull, 2006; Tychanska, 2009; Verderosa, 2007) were conducted in the same urban area in the northeast and most are unpublished dissertations conducted at the same university (Aziz, 2009; Dhaniram-Beharry, 2008; Lella Souravlis, 2010; Nieves-Brull, 2006; Tychanska, 2009;

Verderosa, 2007). Geographically restricted samples limit the significance and generalizability of results to the population of ELLs in other regions throughout the U.S.

The cultural characteristics of the participants are a final concern about sampling. Some C-LIM studies (Brown, 2008; Nieves-Brull, 2006; Templeton, 2012; Verderosa, 2007) have primarily focused on individuals who identified themselves as Hispanic and spoke Spanish as their first language, or other studies (Aziz, 2009; Kranzler et al., 2010; Lella Souravlis, 2010, Tychanska, 2009, Styck, 2012) grouped together individuals from different racial and linguistic backgrounds. Only one study (Dhaniram-Beharry, 2008) conducted analyses with cases separated by race (i.e., Caucasian, Hispanic, and African American) and found that the results did not vary as a function of race.

Measures

Measures of acculturation and language proficiency have not been used in many of the studies (Aziz, 2009; Lella Souravlis, 2010; Nieves-Brull, 2006; Tychanska, 2009, Styck, 2012). Lack of data about these variables limits the extent to which the C-LIM interpretation guidelines developed by Flanagan and colleagues (2007) can be tested. The absence of such measures is a limitation in using an archival dataset; the information is not always available or accessible. In one study (Verderosa, 2007), measures of language proficiency and acculturation were gathered, but were measured through parent report.

Some studies (Dhaniram-Beharry, 2008; Nieves-Brull, 2006; Tychanska, 2009; Verderosa, 2007) used cognitive measures that are dated, such as the WISC-III and DAS. Constructs assessed in cognitive measures often change between revisions; thus, it is unclear how the results obtained for a dated measure would compare to those obtained for the current version. In addition, issues such as the Flynn effect (1999) can influence the comparability of

scores across test batteries.

C-LIM Interpretation

Several authors (Aziz, 2009; Brown, 2008; Dhaniram-Beharry, 2008) used a single comparison score in their analyses rather than the range of possible scores for each cell as recommended by Flanagan and colleagues (Flanagan & Ortiz, 2001; Flanagan et al., 2007). Thus, it is unclear where the predicted values came from. Furthermore, using Flanagan and colleagues' guidelines for C-LIM ranges (see Figure 4, p. 12) in the statistical analyses instead of a single score might have resulted in different findings. For example, Brown (2008) used a predicted score for level 5 of the C-LIM that did not match the scores Flanagan and colleagues had provided. Brown's findings would have been different—in favor of the C-LIM—if Flanagan and colleagues' range had been used instead.

Statistical Analyses

Data management and the statistics used are major limitations of most studies. One, several studies (Aziz, 2009; Kranzler et al., 2010; Lella Souravlis, 2010; Tychanska, 2009, Styck, 2012) grouped individuals from multiple cultural backgrounds (i.e., Lella Souravlis, 2010; Tychanska, 2009) and languages (i.e., Aziz, 2009; Kranzler et al., 2010; Lella Souravlis, 2010; Styck, 2012; Tychanska, 2009) together for analyses. As it is unknown whether differences in C-LIM patterns across cultural and linguistic groups exist, grouping individuals from multiple different cultural and language backgrounds muddles findings and hinders generalization of results from one study to another.

Two, several studies (Aziz, 2009; Dhaniram-Beharry, 2008; Tychanska, 2009) grouped data from multiple test batteries (i.e., WISC-III, WISC-IV, and WJ-III for Aziz, 2009), precluding a clear interpretation of results. In addition, the use of multiple test batteries

introduces variation in psychometric properties (e.g., standard error) into analyses as well as differences in construct representation across measures, a concern that has been raised about XBA (Glutting et al., 2003).

Finally, and most importantly, most studies (Aziz, 2009; Brown, 2008; Dhaniram-Beharry, 2008; Lella Souravlis, 2010; Tychanska, 2009; Verderosa, 2007, Cormier, 2012, Templeton, 2012), only conducted group-level statistics. There were three exceptions. Kranzler and colleagues (2010) and Nieves Brull (2006) conducted descriptive statistics, albeit only frequencies, related to the presence of the C-LIM pattern on an individual level, and Styck (2012) conducted diagnostic accuracy statistics to investigate the utility of the C-LIM for individual use. Utilization of group-level statistics is inconsistent with the guidelines (Flanagan et al., 2007) for interpretation of the C-LIM, which is supposed to be made at an individual level. Aggregating individual data for group analyses misrepresents what the findings may have been if scores were entered into the matrix separately, and interpretation was made on an individual basis. It does not make sense to evaluate a method that is designed for use at an individual level with statistics that are meant to identify group differences. Furthermore, the presence of statistically significant group differences does not necessarily mean that a test can discriminate between individuals holding different group membership (e.g., non-English speakers versus English speakers or bilingual Spanish-English speakers [proficient in both languages] versus proficient Spanish speakers/English language learners; Watkins, Glutting, & Youngstrom, 2005).

Rationale and Research Questions for Present Study

School psychologists are responsible for conducting fair and equitable assessments for all students. Concerns regarding the cultural loading and linguistic demand of commonly used cognitive assessments along with the scarcity of bilingual school psychologists influence the methods school psychologists use to conduct appropriate evaluations of the cognitive abilities of

ELLs. Commonly used approaches for psychological assessment of ELLs (e.g., use of nonverbal cognitive measures and interpreters) have limitations that can negatively influence the quality of the evaluation and the potential interpretations and outcomes associated with such evaluations.

The Culture-Language Interpretive Matrix (C-LIM; (Flanagan & Ortiz, 2001; Flanagan et al., 2007), an extension of XBA, is designed to address the need for a systematic method that takes into account the cultural and linguistic influences of cognitive assessments. Interpretation of the C-LIM is advocated for in several practitioner-oriented publications (Flanagan et al., 2007; Rhodes, Ochoa, & Ortiz, 2005) and is currently being used in elementary and secondary schools in the U.S. to determine the validity of test scores in evaluations for ELLs. As a result, C-LIM interpretation is contributing to high stakes decisions, such as identification of disabilities and eligibility for special education services. It is of concern that the C-LIM is being used within the realm of psychological practice, when minimal research has been conducted on the matrix and findings are unclear at best. Further research on the C-LIM is necessary to address the limitations and gaps in previous research.

The purpose of the present study was to investigate the diagnostic utility of the C-LIM with the WJ-III-NU in differentiating between Spanish-speaking ELLs and monolingual, English-speaking students. Archived cognitive test scores were gathered from the files of ELLs and monolingual English speakers between the ages of 7 and 18 who completed a comprehensive evaluation to determine eligibility for special education services. The following research question was addressed: Based upon Flanagan and colleagues' (2007) predicted range of scores for ELLs, how accurately will the C-LIM scoring pattern discriminate between Spanish-speaking, ELLs and monolingual, English speakers?

METHOD

Participants

Archival data were obtained from the school records of students in a school district in Northern Virginia. These students had been referred for an initial evaluation to determine special education eligibility between 2007 and 2013. Several criteria were used to determine the inclusion or exclusion of cases in the sample. All students had to have completed 14 Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update (WJ-III-NU) subtests, which are associated with the seven broad ability areas typically assessed for referrals in which a learning disability is suspected, and also contribute to an extended General Intellectual Ability (GIA) score. For the monolingual English-speaking (MES) group, students were excluded if a language other than English was spoken in the home. All English language learners (ELLs) spoke Spanish as their first language. Furthermore, cases were included in the ELL group if (a) a standardized English language proficiency score was available (administered within a year of the WJ-III-NU administration), (b) the student was identified as an ELL but was exited from services more than a year before the psychological evaluation, or (c) Spanish was spoken in the home, but the student did not qualify for ELL services. Re-evaluations were excluded as the cases of interest for this study were students who had not previously been evaluated for special education services.

Based on the above criteria, the initial (unmatched) sample consisted of 265 students, representing two different subsamples: (a) 81 Spanish-speaking, ELLs (31%) and (b) 184 MES students (69%). At the time of evaluation, students ranged in age from 6 years, 2 months to 17 years, 7 months (M = 9 years, 9 months; SD = 2 years, 3 months). Ninety-six percent of the MES group and 88 percent of the ELL group were born in the U.S. A summary of demographic information for the sample is presented in Table 2.

Table 2

Demographic Characteristics of the Initial Sample based on Percentage

Characteristic	ELL (n = 81)	MES (n = 184)	Total (n = 265)
	%	%	%
Gender			
Male	55.6	60.3	58.9
Female	44.4	39.7	41.1
Race/Ethnicity[a]			
White	3.7	72.8	51.7
Black	1.2	10.4	7.5
Hispanic, Any Race	93.9	7.6	34.0
Asian	0.0	2.7	1.9
American Indian	0.0	0.0	0.0
Multi-Racial	1.2	6.5	4.9
Age at Time of Evaluation			
6 – 8 Years	46.9	45.7	46.0
9 – 12 Years	50.6	37.5	41.5
13 – 18 Years	2.5	16.8	12.5
Grade at Time of Evaluation			
K – 2	34.6	33.2	33.6
3 – 5	56.8	43.5	47.5
6 – 8	7.4	18.4	15.1
9 – 12	1.2	4.9	3.8
Free/Reduced Lunch Eligibility			
Free Eligible	58.0	10.9	25.3
Reduced Eligible	5.0	1.6	2.6
Not Eligible[b]	37.0	87.5	72.1
Birth Country			
United States	87.8	96.3	93.5
El Salvador	2.5	0.0	0.7
Japan	0.0	1.2	0.7
Mexico	2.5	0.0	0.7
Colombia	0.0	0.5	0.4
Dominican Republic	1.2	0.0	0.4
Guatemala	0.0	0.5	0.4
Honduras	1.2	0.0	0.4
Kazakhstan	0.0	0.5	0.4
Nicaragua	1.2	0.0	0.4
Peru	1.2	0.0	0.4
Puerto Rico	1.2	0.0	0.4
Russia	0.0	0.5	0.4
Venezuela	1.2	0.0	0.4
Unavailable	0.0	0.5	0.4

Table 2 (continued)

Characteristic	ELL (n = 81)	MES (n = 184)	Total (n = 265)
	%	%	%
General Referral Concern[c]			
Academic	95.1	79.9	84.5
Behavioral	28.4	41.8	37.7
Socio-emotional	2.5	14.7	10.9
Eligibility for Special Education			
Eligible for Services	77.8	67.9	70.9
Not eligible for Services	19.8	28.3	25.7
Graduated/Inactive	2.4	3.8	3.4
Disability Category[c]			
Specific Learning Disability (SLD)	50.6	33.7	38.9
Speech or Language Impairment (SLI)	21.0	1.6	7.5
Other Health Impairment (OHI)	18.5	22.3	21.1
Emotional Disability (ED)	2.5	7.1	5.7
Autism (AUT)	1.2	3.3	2.6
Intellectual Disability (ID)	1.2	0.0	0.4
Hearing Impairment (HI)	1.2	0.0	0.4
504 Plan Eligible	1.2	2.2	1.9

Note. ELL = English language learner, MES = monolingual English speaker. [a] Race and ethnicity have been grouped together for ease of presentation. The school district also collects information on race and ethnicity as separate categories in accordance with federal guidelines. [b] Includes cases in which the student's account is no longer active in the district database. [c] Some students had referral concerns listed under more than one category (e.g., Academic and Behavioral) and were found eligible for special education services under more than one category (e.g., SLD and OHI).

An attempt was made to proportionally match the ELL and MES groups based on (a) age, (b) gender, (c) GIA score, (d) free or reduced lunch eligibility and (e) general referral concern. The objective of the matching process was to reduce sources of bias in the sample while simultaneously retaining the greatest number of cases as possible. Initially, cases were matched using a 1:2 (ELL: MES) ratio on a minimum of three (i.e., age, gender, and GIA score) or more variables. Cases were identified as matches if the age, gender, and overall cognitive ability level (e.g., Low Average, Average, High Average) were the same. For all matches, students born within 12 months of each other were considered the same age (i.e., a 12 year, 3 month old could

match with a 13 year, 1 month old as they are 10 months apart). Using this criteria, 36 percent (*n* = 96) of the total sample (*N* = 265) were matched. Due to the poor retention of cases during the initial matching procedure, matches based on free or reduced lunch eligibility and general referral concern were not attempted as even fewer cases would be retained.

A second matching procedure was completed with cases matched by age and gender using a 1:2 (ELL: MES) ratio. Based on this second procedure, 88 percent (*n* = 234) of the total sample was matched. Given that the purpose of matching was to minimize sources of bias while retaining the greatest number of cases, it was clear that matching based on age (years and months) and gender using a 1:2 (ELL: MES) ratio best met this objective. Two sources of bias were reduced and 88 percent of the sample was retained. Thus, the matched sample (*n* = 234) was created with subsamples that consisted of 78 Spanish-speaking, ELLs (33%) and 156 MES students (67%). At the time of the evaluation, students in the matched sample ranged in age from 6 years, 2 months to 15 years, 7 months (*M* = 9 years, 2 months; SD = 1 year, 10 months). Ninety-six percent of the MES group and 90 percent of the ELL group were born in the U.S. A summary of demographic information for the matched sample is presented in Table 3.

In regard to language, parents reported that all of the ELLs were exposed to Spanish at home and two ELLs were exposed to multiple languages in the home (Assyrian or Farsi in addition to Spanish and English). Approximately 74 percent of the ELLs in the initial (unmatched) sample preferred English, 25 percent preferred English and Spanish equally, and 1 percent preferred Spanish. A majority of the ELLs in the unmatched sample (98%) was educated entirely in the U.S. On entering the district, 90 percent of ELLs in the unmatched sample were screened for language services, with 83 percent receiving such services at some point during their education. Furthermore, 70 percent of the ELLs in the unmatched sample received

Table 3

Demographic Characteristics of the Matched Sample based on Percentage

Characteristic	ELL (*n* = 78)	MES (*n* = 156)	Total (*n* = 234)
	%	%	%
Gender			
Male	57.7	57.7	57.7
Female	42.3	42.3	42.3
Race/Ethnicity[a]			
White	3.8	75.0	51.3
Black	1.3	8.2	6.0
Hispanic, Any Race	93.6	7.1	35.9
Asian	0.0	2.6	1.7
American Indian	0.0	0.0	0.0
Multi-Racial	1.3	7.1	5.1
Age at Time of Evaluation			
6 – 8 Years	48.7	53.8	52.1
9 – 12 Years	48.7	42.4	44.5
13 – 18 Years	2.6	3.8	3.4
Grade at Time of Evaluation			
K – 2	35.9	39.1	38.0
3 – 5	55.1	50.6	52.1
6 – 8	7.7	9.0	8.6
9 – 12	1.3	1.3	1.3
Free/Reduced Lunch Eligibility			
Free Eligible	57.7	10.3	26.1
Reduced Eligible	5.1	0.6	2.1
Not Eligible[b]	37.2	89.1	71.8
Birth Country			
United States	89.6	95.7	93.8
El Salvador	1.3	0.0	0.4
Japan	0.0	1.3	1.0
Mexico	1.3	0.0	0.4
Colombia	0.0	0.6	0.4
Dominican Republic	1.3	0.0	0.4
Guatemala	0.0	0.6	0.4
Honduras	1.3	0.0	0.4
Kazakhstan	0.0	0.6	0.4
Nicaragua	1.3	0.0	0.4
Peru	1.3	0.0	0.4
Puerto Rico	1.3	0.0	0.4
Russia	0.0	0.6	0.4
Venezuela	1.3	0.0	0.4
Unavailable	0.0	0.6	0.4

Table 3 (continued)

Characteristic	ELL (*n* = 78)	MES (*n* = 156)	Total (*n* = 234)
	%	%	%
General Referral Concern[c]			
Academic	96.2	82.1	86.8
Behavioral	28.2	41.0	36.8
Socio-emotional	1.3	13.5	9.4
Eligibility for Special Education			
Eligible for Services	78.2	68.6	71.8
Not eligible for Services	19.2	30.1	26.5
Graduated/Inactive	2.6	1.3	1.7
Disability Category[c]			
Specific Learning Disability	52.6	34.6	40.6
Speech or Language Impairment	20.5	1.9	8.1
Other Health Impairment	17.9	23.7	21.8
Emotional Disability	2.6	5.1	4.3
Autism	1.3	3.2	2.6
Intellectual Disability	1.3	0.0	0.4
Hearing Impairment	1.3	0.0	0.4
504 Plan Eligible	1.3	2.6	2.1

Note. ELL = English language learner, MES = monolingual English speaker. [a]Race and ethnicity have been grouped together for ease of presentation. The school district also collects information on race and ethnicity as separate categories in accordance with federal guidelines. [b]The "Not Eligible" lunch status includes cases in which the student's account is no longer active in the district database. [c]Some students had referral concerns listed under more than one category (e.g., Academic and Behavioral) and were found eligible for special education services under more than one category (e.g., SLD and OHI).

language services at the time of the multidisciplinary evaluation. Language proficiency levels and scaled scores were available for 72 percent of the ELL group and are presented for the matched sample and the initial sample in Table 4. During the psychological evaluations from which data were pulled for this study, psychologists interpreted approximately 78 percent of the ELL cases using the C-LIM. Findings for the matched sample were identical or differed by only one percentage point to the initial sample.

Measures

The following information was obtained from the school district for all students in the

Table 4

Classification of English Language Proficiency Data from the ACCESS for ELLs Test for ELLs in Matched and Unmatched Samples (n = 58)

Proficiency Level/Score	Unmatched Sample	Matched Sample
	%	%
Expected Level/Tier		
Level KG[a]	3.4	3.6
Tier A – Levels 1 – 3	17.2	17.9
Tier B – Levels 2 – 4	58.6	57.1
Tier C – Levels 3 – 5	20.8	21.4
Scale Scores		
100 – 199	1.7	1.8
200 – 299	27.6	28.6
300 – 399	70.7	69.6
400 – 499	0.0	0.0
500 – 600	0.0	0.0
Overall Proficiency Level		
1.0 – 1.9	3.4	3.6
2.0 – 2.9	8.6	8.9
3.0 – 3.9	57.0	55.4
4.0 – 4.9	22.4	23.2
5.0 – 6.0	8.6	8.9

Note. ACCESS for ELLs = Assessing Comprehension and Communication in English State-to-State for English Language Learners.
[a] If no information is available to place a student in a tier based on their expected language proficiency, Level KG is administered; it is not subject to the limits of the tier system.

sample: (a) ELL status (yes/no) (b) country of birth, (c) age, (d) grade, (e) race/ethnicity, (f) gender, (g) eligibility for free or reduced lunch, (h) reason for special education referral, (i) date of WJ-III-NU evaluation; and (j) WJ-III-NU test, cluster, and GIA scores. In addition, the following information was gathered for students identified as ELLs: (a) languages spoken or understood by the student, (b) language(s) preferred by the student, (c) languages spoken in the home, (d) time spent in schools outside the U.S., (e) U.S. entry date, (f) U.S. entry age, (g) time spent in U.S. schools, (h) screened for ELL services (yes/no), (i) current ELL status, (j) time

spent in ELL program, (k), English proficiency level (scores from within a year of WJ-III-NU assessment date), and (l) C-LIM used within psychological evaluation (yes/no).

Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update (WJ-III-NU)

The WJ-III-NU (Woodcock et al., 2007) is an individually administered measure of intelligence that provides a broad assessment of overall cognitive ability for individuals from 2 to 90 years of age and older. The Cattell-Horn-Carroll (CHC) theory of intelligence was used to guide the development of the WJ-III-NU (Schrank, McGrew & Woodcock, 2001). Seven broad abilities associated with academic skills are assessed within the WJ-III-NU: (a) crystallized intelligence, (b) fluid reasoning, (c) long-term retrieval, (d) visual-spatial thinking, (e) auditory processing, (f) processing speed, and (g) short-term memory. Measurement of two narrow abilities subsumed within each broad ability contributes to a cluster score for each broad ability and as well as an overall General Intellectual Ability (GIA) score. Scores from the WJ-III-NU are expressed as standard scores (M = 100; SD = 15). The 14 tests that contribute to each of the 7 broad ability cluster scores are listed in Table 5.

The norm sample for the WJ-III-NU consisted of a nationally representative sample of 8,818 individuals, who were randomly selected and stratified based on variables, including, but not limited to, geographic region, sex, race, and Hispanic (or non-Hispanic) background. School-age (kindergarten through 12th grade) norms for the WJ-III-NU are based on a sample of 4,783 individuals. Census projections for the year 2000 were used to develop initial norms for the WJ-III, but were revised and published as the WJ-III-NU after final census data for 2000 were released.

Reliability of the scores for WJ-III-NU tests was calculated either through split-half

Table 5

Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update Tests

Cluster	Test
Crystallized Intelligence	Verbal Comprehension General Information
Fluid Reasoning	Concept Formation Analysis-Synthesis
Long-Term Retrieval	Visual-Auditory Learning Retrieval Fluency
Visual-Spatial Thinking	Spatial Relations Picture Recognition
Auditory Processing	Sound Blending Auditory Attention
Processing Speed	Visual Matching Decision Speed
Short-Term Memory	Numbers Reversed Memory for Words

procedures with the Spearman-Brown correction formula or test-retest procedures (for speeded tests or tests with multiple point items). Median reliabilities for test scores within the WJ-III-NU ranged from .74 to .97, with only two values falling below .80. Median reliabilities for the scores of broad ability clusters were above .80, and all but three exceeded .90.

Validity evidence for the WJ-III-NU was gathered through examining the content of the tests in aligning with CHC theory, developmental evidence based upon growth curves across age, evidence of factor structure through confirmatory factor analysis (CFA; conducted on the WJ-III standardization data), and concurrent validity based upon correlations with four other commonly used measures of intelligence. CFAs were conducted using two models that were expected to provide the best fit based on CHC theory: (a) *g* and 9 broad CHC abilities; and (b) *g*, 9 broad CHC abilities, and 17 narrow CHC abilities. Each CFA was initially conducted with

individuals at age 6 and older, but the *g* and broad ability model was also tested across five age groups spanning the entirety of the ages represented in the standardization sample. Three types of fit statistics were used in which smaller values of chi square and Akaike information criterion (AIC) were associated with the better fitting model, and root mean square error of approximation (RMSEA; Byrne, 2001) was equal or less than .05. Results across the entire age range indicated that the broad CHC ability model provided the best fit to the data (χ^2 (536) = 13,189.16, AIC = 13,377.16, RMSEA = .056) compared to the six other models tested. When separated into five age groups, the broad CHC ability model, again, provided the best fit (χ^2 (536) = 3,221.86 – 5, 557.34, AIC = 3, 409.86 – 5,745.34, RMSEA = .058 – .073). In addition, results for the second CHC model (broad and narrow abilities) across ages provided a good fit to the WJ-III standardization data (χ^2 (1,115) = 22,348.80, RMSEA = .050). The AIC was not reported. Median correlations between the WJ-III-NU Extended GIA score and the full scale or composite scores of the Differential Ability Scales (DAS), Wechsler Preschool and Primary Scale of Intelligence – Revised (WPPSI-R), Stanford-Binet Intelligence Scale – Fourth Edition (SB-IV), and Wechsler Intelligence Scale for Children – Third Edition (WISC-III) ranged from .71 to .76. Detailed information regarding the psychometric properties of the WJ-III-NU is provided in the technical manual (McGrew, Schrank, & Woodcock, 2007).

Independent investigations into the validity of WJ-III scores have also been conducted. Two studies examined the factor structure of the WJ-III or WJ-III-NU across age. Taub and McGrew (2004) examined via CFA the factor structure of 14 of the 20 available tests (i.e., all of the tests listed in Figure 1) contained in the WJ-III standardization sample (*n* = 7,485) across five age groups (individuals ranging from 6 to 90+ years of age). Two primary fit criteria were used: (a) goodness of fit index (GFI; $\geq$ .90 = adequate fit; $\geq$ .95 = excellent fit; Hu & Bentler, 1999)

and (b) RMSEA ≤ .05 = good fit; Byrne, 2001). Results indicated that the expected seven-factor model was invariant across all five age groups; for efficiency in presentation, however, only standardized values averaged by age group were provided (GFI = .958, RMSEA = .025). Floyd, McGrew, Barry, Rafael, and Rogers (2009) examined *g* loadings and specificity estimates for each of seven factor clusters (Gc, Clr, Gv, Ga, Gf, Gs, and Gsm) using data from the WJ-III-NU standardization data (*n* = 3,577, divided into seven age-based samples and ranging in age from 4 to 60+ years). Floyd et al. examined if any of the seven factors demonstrated high specificity effects, which could support independent interpretation of factor scores beyond what would be offered by interpretation of *g*. Effect sizes greater than .70 were considered significant. Floyd et al. found that Gc, Glr, and Gf seemed to be primarily measures of a general factor across a majority of age levels, whereas Gv, Ga, and Gs appeared to primarily measure specific abilities across most ages. Findings were mixed for Gsm (with high specificity effects at only two levels and general effects at others).

Several researchers have also examined the Woodcock-Johnson factor structure across ethnic groups. Keith (1999) examined the effects of general and specific cognitive abilities on achievement for African American, Hispanic, and Caucasian groups. Although this study was conducted using an outdated measure (the WJ-R), this was the only study available that examined the WJ factor structure for Hispanic individuals. Multi-sample structural equation modeling was used to examine if the same specific cognitive abilities were important for reading and math achievement across ethnic groups. Separate models were calculated for three grade level groupings (1-4, 5-8, and 9-12). The comparative fit index (CFI) was used to evaluate model fit (≥ .90 = adequate fit; ≥ .95 = excellent fit; Hu & Bentler, 1999). Keith found that specific abilities were important in predicting math and reading achievement across grade level

groups beyond that which could be due to *g*. In addition, the same specific cognitive abilities were found to be invariant across African American, Hispanic, and Caucasian groups. The magnitude of the influence of specific cognitive abilities on reading and math achievement was also similar across ethnic groups, with only two exceptions. Gc and Gs demonstrated statistically similar effects on Passage Comprehension for all students in grades five through eight when paths were constrained. These two abilities were found to be more important for the Hispanic group as the CFA model fit better when paths to these specific abilities were allowed to vary than when paths for all three groups were invariant (CFI = .957, RMSEA = .025-.040, $\Delta\chi^2$ = 12.670[2], p = .002). Keith concluded that Gc and Gs may be more important for Hispanic students' reading comprehension skills due to bilingualism. More recently, Edwards and Oakland (2006) used CFA on WJ-III standardization data to examine the factor structure of the WJ-III for African Americans and Caucasian Americans. The factor structure of the WJ-III was invariant across African American and Caucasian American groups (CFI = .99, RMSEA = .059), but the two groups differed on mean scores, with African Americans consistently scoring lower across all WJ-III tests. A search for studies investigating the validity of WJ-III scores for ELLs yielded no results.

Joint CFAs have also been completed on the WJ-III paired with the Cognitive Assessment System (CAS; Keith, Kranzler, & Flanagan, 2001), WISC-III (Phelps, McGrew, Knopik, & Ford, 2005), DAS (Sanders, McIntosh, Dunham, Rothlisberg, & Finch, 2007), and Delis-Kaplan Executive Function System (D-KEFS; Floyd, Bergeron, Hamilton, & Parra, 2010). Keith et al. and Floyd et al. used independent samples (n = 155 and n = 100, respectively), while Phelps et al. and Sanders et al. used the WJ-III standardization sample. In each of these studies, the WJ-III tests generally loaded on the expected CHC factors as outlined in Table 4. One

notable exception was reported; Floyd et al. found that the Analysis-Synthesis and Concept Formation tests from the WJ-III loaded on the Gc and Executive Function factors, respectively, instead of loading together on the hypothesized Gf factor.

Criterion-related validity evidence has also been gathered to demonstrate the relation between CHC cognitive abilities in the WJ-III Tests of Cognitive Abilities and achievement in writing (Floyd, McGrew, & Evans, 2008), math and reading (McGrew & Wendling, 2010) on the WJ-III Tests of Achievement. Findings were mixed across studies, with some specific cognitive abilities demonstrating moderate to strong effects on achievement (e.g., the moderate to strong effect between Gc and basic reading skills), while others (e.g., Gv) were found to have negligible effects on reading, writing, and math achievement.

Assessing Comprehension and Communication in English State-to-State for English Language Learners (ACCESS for ELLs)

The World-Class Instructional Design and Assessment (WIDA; 2011) Consortium and Center for Applied Linguistics (CAL; 2012) developed ACCESS for ELLs to assess English language proficiency skills in students in kindergarten through 12th grades who have been identified as ELLs (Kenyon, 2006). ACCESS for ELLs aligns with Virginia's state standards (and standards for 23 other states) for English language proficiency put forth for ELLs as well as state standards for academic content (i.e., language arts, social studies, science, and math). It contains evaluation of social and academic proficiency in English across listening, speaking, reading, and writing domains. ELLs' proficiency in English is classified into one of six levels: (1) Entering, (2) Beginning, (3) Developing, (4) Expanding, (5) Bridging, or (6) Reaching. The six levels of language proficiency are considered to reflect a continuum of language development with the first level (Entering) signifying the lowest level of English language proficiency (e.g.,

understanding of pictures, graphics, and single words) and the sixth level (Reaching) signifying that a student has reached a level of proficiency in English such that the student should be successful in an English-only, regular education classroom without external support. Proficiency level scores on the ACCESS for ELLs range from 1.0 to 6.0, with 3.5 as a center point. The proficiency level scores associated with the test do not coincide with points earned, but rather the level of latent English language proficiency the student has demonstrated. Group administration is used for the listening, reading, and writing sections of ACCESS for ELLs, and responses for each of these areas are either machine scored (listening and reading) or scored by trained individuals (writing) other than the examiner. The speaking portion of the ACCESS for ELLs is administered individually and scored by the examiner during administration.

Development of the ACCESS for ELLs test involved preparation and review of items and tasks as well as pilot testing (N = 1,314 students in grades K-12 across three states) and field testing (N = 6,662 students in grades K-12 across eight states). Ninety-six native languages other than English were represented in the field testing sample. Rasch reliability indexes of the scores were calculated for listening (r = .82) and reading (r = .91) tasks, while interrater reliability of the scores was calculated for writing (r = .97) and speaking (r = .90 to .92 across grade levels) tasks. Concurrent validity evidence was gathered through examination of the relationship between ACCESS for ELLs and the Language Assessment Scales (LAS; DeAvila & Duncan, 1977), IDEA Proficiency Test (IPT; Ballard, Tighe, & Dalton, 1991), Language Proficiency Test Series (LPTS; Plake, Impara, & Spies, 2003), and Maculaitis II Test of English Language Proficiency (MACII; Maculaitis, 2003) test. A sample of 4,985 students took one of the four other language proficiency measures two months prior to taking ACCESS for ELLs. Correlations between ACCESS for ELLs and the LAS, IPT, LPTS, and MACII were generally in

the moderate to high range across listening (r = .47 - .61), speaking (r = .51 - .66), reading (r = .58 - .77), and writing (r = .55 - .71) domains. Content validity of the listening and reading items of the ACCESS for ELLs test was also studied using more than 6,500 students in grades K-12 from the field test sample. It was found that for both listening and reading domains increases in average item difficulty corresponded with increases in language proficiency for all proficiency levels. The only exception to this finding was for levels four (Expanding) and five (Bridging) in reading where item difficulty did not differentiate between individuals classified at levels four and five. The structural validity of the ACCESS for ELLs test has not been examined. Additional technical reports are created on an annual basis to add to the validity evidence for the ACCESS for ELLs and are available on the WIDA Consortium website (Kenyon, 2006). To date, no external investigations of the psychometric characteristics of ACCESS for ELLs could be found.

Acculturation Level

Years of residency in the U.S. at the time of evaluation were used to measure level of acculturation as no formal acculturation assessment is given in the school district in which data were obtained. Criteria for level of acculturation were operationalized using the guidelines Flanagan et al. (2007, p. 200) provided. Specifically, students who lived in the U.S. for greater than seven years were considered highly acculturated (consistent with criterion for *Slightly Different*), students who lived in the U.S. for three to seven years were considered moderately acculturated (consistent with criterion for *Moderately Different*), and students who lived in the U.S. for less than three years were considered slightly acculturated (consistent with criterion for *Markedly Different*).

Procedure

The school district had collected data about their students, who were the focus of this study, at various district events, such as school registration (home language survey, demographic information), referrals and evaluations for special education eligibility (dual language assessment, psychological evaluation), and determination of qualification for ELL services (language screening and proficiency scores, dates of service). Permission to access extant student data was obtained from the research office of the school district in which the data were gathered. Data were obtained from two sources: (a) the school psychologists and (b) student files. Once permission was granted, school psychologists employed in the district were contacted via their supervisor to obtain psychological evaluation reports for students who had been evaluated using the WJ-III-NU. Psychologists searched their computer files for applicable reports and sent them to the district's supervisor of psychological services. At this point, all identifying information was removed from the documents and a research identification number was assigned to each report. On obtaining the de-identified reports, the principal investigator saved them and entered relevant information into a database for the study. Then, the district's research office was contacted via the psychologists' supervisor to obtain information not contained in the psychological reports. The research office gathered the additional information for relevant cases from the district database and sent it to the principal investigator using assigned research identification numbers.

Data Management and Analyses

The design for this study was non-experimental. The Statistical Package for the Social Sciences, version 17.0 (SPSS 17.0) was used to conduct all statistical analyses. A series of steps were taken to address the research question in this study. First, descriptive statistics, including means, standard deviations, and ranges, were calculated for WJ-III-NU test, cluster, and GIA

scores. Second, each case was classified as following or not following the expected C-LIM pattern based on three definitions (or levels) of a declining pattern of scores (i.e., most stringent interpretation, moderately stringent interpretation, least stringent interpretation). Third, diagnostic utility statistics and receiver operating curve (ROC) analyses were conducted across the three levels of interpretation. Fourth, follow-up analyses were conducted to determine whether the findings from diagnostic utility statistics and ROC were associated with specific patterns as delineated by Flanagan et al. (2013). What follows is a description of the classification process used to delineate whether ELL and MES cases followed the expected C-LIM pattern and primary statistics used: diagnostic utility statistics and ROC analyses.

Most Stringent C-LIM Interpretation

Cases were evaluated based on the most stringent criteria because this interpretation is consistent with guidelines offered by Flanagan, Ortiz, and Alfonso (2007, 2013). For the most stringent C-LIM interpretation, test scores from all nine cells in the C-LIM were used to interpret the decline in performance. For ELLs, declines across all cells were expected. However, the order in which cells in the C-LIM were inspected differed slightly depending on the type of influence on the student's scores (i.e., cultural loading only, linguistic demand only, or the combined influence of cultural loading and linguistic demand).

For the singular influence of cultural loading, the expected pattern of decline in an ELL's performance was based primarily on level of cultural loading (low, moderate, or high) and secondarily on level of linguistic demand. Thus, the order of interpretation for cultural loading was as follows: (a) low cultural loading/low linguistic demand, (b) low cultural loading/moderate linguistic demand, (c) low cultural loading/high linguistic demand, (d) moderate cultural loading/low linguistic demand, (e) moderate cultural loading/moderate linguistic demand, (f)

moderate cultural loading/high linguistic demand, (g) high cultural loading/low linguistic demand, (h) high cultural loading/moderate linguistic demand, (i) high cultural loading/high linguistic demand. The expected pattern of decline for the singular influence of linguistic demand on an ELLs performance was based primarily on level of linguistic demand (low, moderate, or high), and secondarily on level of cultural loading. The order of interpretation for linguistic demand was as follows: (a) low linguistic demand/low cultural loading, (b) low linguistic demand/moderate cultural loading, (c) low linguistic demand/high cultural loading, (d) moderate linguistic demand/low cultural loading, (e) moderate linguistic demand/moderate cultural loading, (f) moderate linguistic demand/high cultural loading, (g) high linguistic demand/low cultural loading, (h) high linguistic demand/moderate cultural loading, (i) high linguistic demand/high cultural loading. Separate graphs depicting the singular influences of cultural loading or linguistic demand are included in the software distributed along with the *Essentials of Cross-Battery Assessment, Third Edition* (Flanagan et al., 2013). The expected pattern of decline for an ELL based on the combined influence of cultural loading and linguistic demand was presented in Figure 3 (see p. 11), and is also included in the software (Flanagan et al., 2013).

For this study, a case was considered to follow the declining pattern according to most stringent criteria if the average cell scores decreased across all nine cells in the C-LIM in the order explained above. Thus, three potential declining patterns following the most stringent criteria were evaluated: (a) a decline in cultural loading only, (b) a decline in linguistic demand only, and (c) a combined decline in cultural loading and linguistic demand. Each case was classified as to whether the scores followed or did not follow the expected pattern (for a, b, and c patterns), and each case was dichotomously coded to reflect that determination (0 = no, 1 = yes).

Moderately Stringent C-LIM Interpretation

Cases were also evaluated using a moderately stringent criterion because it allows for students to have some variability in test scores in the C-LIM and still follow a declining pattern. For interpretation of the declining test scores in the C-LIM, test scores from all nine cells must be included in the calculation of the pattern of decline, but is based on the five level system discussed earlier (see Figure 5, p. 13). The five-level system has been used in previous research (Brown, 2008; Dhaniram-Beharry, 2008; Lella Souravlis, 2010).

To use this five level system of interpretation, average scores were created for the combined decline across cultural loading and linguistic demand. Cases that demonstrated a decrease in scores across all five levels were designated as following the expected pattern of decline. To consider singular declines in either cultural loading or linguistic demand, averages were calculated for low, moderate, and high levels for each classification. Then, the averages were compared and if a decline was evident from low to moderate to high levels, the case was identified as following the expected pattern. If a decline in average scores was not evident from low to moderate to high levels, the case was designated as not following the expected pattern. Again, all cases were dichotomously coded to reflect the determination of whether they followed the expected pattern (0 = no, 1 = yes) for each possible interpretation (i.e., combined decline, cultural decline, or linguistic decline).

Least Stringent C-LIM Interpretation

The third C-LIM interpretation method was determined as the least restrictive option in identifying a declining pattern of scores within the matrix. This option involves a visual examination of the cell averages in the C-LIM (not the additional bar graphs), and thus, is probably the one most frequently used in practice. The interpretation of the pattern of decline in

scores in the C-LIM involves a comparison of averaged test scores from only three cells in the C-LIM for each expected pattern of decline.

Specifically, for this study the combined influence of cultural loading and linguistic demand was assessed through a comparison of the cell average for tests classified with low cultural loading and low linguistic demand (Low) to the cell average for tests with a moderate classification in both cultural loading and linguistic demand (Moderate). Then a comparison was made between the Moderate cell average and the cell average for tests with high cultural loading and high linguistic demand (High). If a decline in scores was observed across each of the three cells (Low to Moderate to High), then the case was identified as following the expected pattern of decline for combined cultural loading and linguistic demand. This pattern of decline is also depicted by the diagonal arrow in Figure 2 (see p. 9). For the singular influence of cultural loading, a comparison was made from top to bottom for the cell averages in the left column of the C-LIM. Interpretation of the influence of linguistic demand involved a comparison of cell averages across the top row of the C-LIM (left to right). These additional patterns are also displayed in Figure 2 (see p. 9). For the least stringent C-LIM interpretation, each case was again dichotomously coded to reflect the determination of whether it followed the expected pattern (0 = no, 1 = yes) for each of the three possible interpretations.

Diagnostic Utility Statistics

To use diagnostic utility statistics, the coded cases based on the interpretation criteria described above were reviewed to determine the number of ELLs and MESs that fell into each classification (i.e., followed pattern or not). Then, diagnostic utility statistics were conducted using the established classifications. Diagnostic utility statistics are used to examine the accuracy of a test in distinguishing between individuals with a diagnosed condition and those

without (Matthey & Petrovski, 2002). The diagnostic utility statistics is based on a matrix of four conditions: (a) true positive, (b) true negative, (c) false negative, and (d) false positive. A true positive (TP) is when a person with a positive test also has the condition. A true negative (TN) is when a person with a negative test does not have the condition. A false positive (FP) is when a person has a positive test result, but does not have the condition. Finally, a false negative (FN) is when a person has a negative test result, but has the condition.

In this study, the relation between a declining pattern (as defined by the most to least stringent guidelines outlined above) in the C-LIM (diagnostic test) and ELL status (target condition) were examined. Thus, a TP was an ELL with cognitive scores that followed a declining pattern, while an FN was an ELL with cognitive scores that did not follow a declining pattern. In addition, a monolingual, English speaking student with scores that followed a declining pattern was considered an FP, while a monolingual, English speaking student with cognitive scores that did not follow a declining pattern was considered a TN. Once all cases were categorized into one of the four classifications (i.e., TP, FN, FP, or TN) for each of the three sets of interpretation guidelines, sensitivity, specificity, positive predictive value, and negative predictive value were calculated. Definitions and associated equations for the diagnostic utility statistics as applied in this study are listed in Table 6 (Henderson, 1993). According to Matthey and Petrovski (2002), the accuracy and economy of a diagnostic test necessitates sensitivity values of at least 0.70 and specificity values of at least 0.80. These guidelines were followed in this study.

Receiver Operating Curve (ROC) Analyses

Diagnostic utility statistics are considered to provide valuable information about the utility of tests, but also have limitations. In particular, sensitivity, specificity, and predictive

Table 6

Definitions and Equations of Diagnostic Utility Statistics

Statistic	Definition	Equation
Sensitivity	Proportion of cases that are ELLs with a declining pattern	TP/(TP + FN)
Specificity	Proportion of cases that are not ELLs with scores not following a declining pattern	TN/(TN +FP)
Positive Predictive Value	Proportion of cases that follow a declining pattern and are correctly identified as ELLs	TP/(TP + FP)
Negative Predictive Value	Proportion of cases that do not follow a declining pattern and are correctly identified as MESs	TN/(TN + FN)

Note. ELL = English language learner, MES = monolingual English speaker, TP = True Positive; FN = False Negative; TN = True Negative; FP = False Positive.

values can change depending on the cutoff values used for the diagnostic test and the prevalence rate of the target condition (McFall & Treat, 1999). To address these limitations, receiver operating curve (ROC) analysis was conducted. ROC analysis involves plotting TP and FP values across all possible cutoff scores (Henderson, 1993). Given that Flanagan et al. (2007, 2013) did not specify a minimum difference between cell averages for the pattern of decline that ELLs are expected to follow, a continuous variable was created that represented the minimum discrepancy between cell averages across an expected pattern of decline. This continuous variable was created for the three options (i.e., combined decline, cultural decline, linguistic decline) under each of the three levels of criteria for C-LIM interpretation (i.e., most stringent, moderately stringent, and least stringent). Once a continuous variable was calculated for each option, the variable was used along with language status (ELL or MES) to plot a curve.

The area under the curve (AUC) value was also calculated. The AUC is an index of accuracy for a diagnostic test in differentiating between individuals with and without the target condition (Henderson, 1993; Swets, 1988). A test is considered to be highly accurate if the AUC value is 0.9 to 1.0, moderately accurate if the AUC is 0.7 to 0.9, and to have low accuracy if the AUC is 0.5 to 0.7 (Swets, 1988). In the context of this study, the AUC represented the degree of accuracy with which the diagonally decreasing C-LIM pattern differentiated between ELLs and MESs.

Acculturation and Language Proficiency Analyses

Flanagan et al. (2007, 2013) asserted that a greater decline performance would be evident in the C-LIM for ELLs with lower levels of acculturation and English language proficiency. In this study, the influences of acculturation and English language proficiency were examined using age of entry into the U.S. and language proficiency test scores, respectively. The relationship between acculturation, language proficiency and the presence of a declining pattern in the C-LIM was investigated. The set of declining patterns specified in the levels of C-LIM interpretation were only expected to occur for ELLs whose cultural and linguistic background, secondary to true cognitive abilities, were the most significant influences on their cognitive test scores (Flanagan et al., 2007, 2013).

RESULTS

The matched sample was used for all primary analyses. First, descriptive statistics and preliminary analyses were conducted on the WJ-III-NU tests. Second, frequency calculations for C-LIM decisions based on most stringent, moderately stringent, and least stringent criteria for interpretation were calculated. Third, comparisons were made between students who followed the patterns of decline and those who did not based on language status (ELL or MES), level of acculturation, and English language proficiency using frequencies and independent samples t tests. Next, diagnostic utility statistics (sensitivity, specificity, ROC curves, and AUC values) were plotted and calculated. The primary diagnostic utility statistics based on variables representing a continuous pattern of decline were conducted first, followed by binary AUC calculations across all possible cut scores for each pattern of decline that were conducted to further investigate the validity of the C-LIM in differentiating between ELLs and MESs. Finally, a set of post hoc analyses (including re-calculations of all of the analyses just described) was completed on a subset of the unmatched sample, namely students not identified with SLD.

Descriptive Statistics

The means and standard deviations for WJ-III-NU tests for the matched ELL and MES samples are contained in Tables 7 and 8, respectively. The descriptive analysis presented is based on the least stringent criteria for C-LIM interpretation (pp. 19-20). In regard to cultural loading (by row), on average, neither the ELL nor the MES sample followed the declining pattern. The same pattern was observed about linguistic demand (by column) for both samples. On average, the ELL sample appeared to follow the declining pattern for the combined influence of cultural loading and linguistic demand, but the MES sample did not.

Preliminary Analyses

A series of independent samples t-tests were conducted to determine whether mean

Table 7

Means and Standard Deviations of WJ-III-NU Subtests for the ELL Matched Sample

Degree of Cultural Loading	Degree of Linguistic Demand: Low		Moderate		High	
	Test Name	***M (SD)***	**Test Name**	***M (SD)***	**Test Name**	***M (SD)***
Low	Spatial Relations	98.8 (8.6)	Numbers Reversed	89.0 (14.9)	Analysis-Synthesis	98.2 (13.1)
			Visual Matching	84.5 (15.6)	Concept Formation	91.6 (12.7)
	Cell Average	**98.8 (8.6)**	**Cell Average**	**86.8 (15.3)**	**Cell Average**	**94.9 (12.9)**
Moderate	Picture Recognition	103.4 (12.0)	Retrieval Fluency	87.8 (16.2)	Auditory Attention	97.4 (15.0)
			Visual Auditory Learning	89.2 (11.2)	Decision Speed	97.2 (16.1)
					Memory for Words	85.9 (13.6)
					Sound Blending	95.9 (12.4)
	Cell Average	**103.4 (12.0)**	**Cell Average**	**88.5 (13.7)**	**Cell Average**	**94.1 (14.3)**
High					General Information	87.7 (15.3)
					Verbal Comprehension	87.8 (10.7)
	Cell Average		**Cell Average**		**Cell Average**	**87.8 (13.0)**

Note. n = 78. WJ-III-NU = Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update; ELL =English language learner.

differences existed between the two language samples (ELL vs. MES) on all of the WJ-III NU tests. Homogeneity of variance was met for all comparisons ($p > .05$). Results of the independent-samples *t* tests indicated that the mean scores for all tests, except one (Decision Speed), were statistically significant. This result was consistent with the expectation that ELLs would generally perform lower than MESs. Effect sizes (Cohen's *d*) ranged from -1.23 to -.16. A summary of the findings is contained in Table 9.

Calculations of Frequencies for C-LIM Decisions

The number of cases that resulted from true positive (TP), false positive (FP), true negative (TN), and false negative (FN) was calculated for C-LIM decisions (decline pattern or no

Table 8

Means and Standard Deviations of WJ-III-NU Subtests for the MES Matched Sample

Degree of Cultural Loading	Degree of Linguistic Demand: Low		Moderate		High	
	Test Name	***M (SD)***	**Test Name**	***M (SD)***	**Test Name**	***M (SD)***
Low	Spatial Relations	102.8 (10.4)	Numbers Reversed	93.1 (12.7)	Analysis-Synthesis	104.2 (13.8)
			Visual Matching	89.9 (14.0)	Concept Formation	104.8 (14.2)
	Cell Average	**102.8 (10.4)**	**Cell Average**	**91.5 (13.4)**	**Cell Average**	**104.5 (14.0)**
Moderate	Picture Recognition	106.3 (10.0)	Retrieval Fluency	97.4 (14.8)	Auditory Attention	104.6 (12.6)
			Visual Auditory Learning	94.6 (14.7)	Decision Speed	99.5 (14.1)
					Memory for Words	99.9 (13.5)
					Sound Blending	107.8 (12.2)
	Cell Average	**106.3 (10.0)**	**Cell Average**	**96.0 (14.8)**	**Cell Average**	**103.0 (13.1)**
High					General Information	105.7 (14.4)
					Verbal Comprehension	101.7 (12.1)
	Cell Average		**Cell Average**		**Cell Average**	**103.7 (13.3)**

Note. n = 156. WJ-III-NU = Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update; MES = monolingual English speaker.

decline pattern) by language group (ELL or MES). A template showing the location of TP, FP, TN, and FN frequencies within the matrix is depicted in Table 10. Results of the frequency calculations are organized based on the level of C-LIM interpretation.

Most Stringent C-LIM Interpretation

For the most stringent interpretation, test scores from all nine cells in the C-LIM were interpreted across the three decline patterns: (a) cultural loading only, (b) linguistic demand only, and (c) combined cultural loading and linguistic demand. An explanation of the specific order of all nine cells in the C-LIM for each of the three influences (a, b, and c) is presented in the Method (pp. 17-18). The combined influence of cultural loading and linguistic demand under

Table 9

Results of Independent Samples t tests on WJ-III-NU Scores Based on Language Status

WJ-III-NU Test	ELL (n = 78)	MES (n = 156)	p value	d
Verbal Comprehension	87.77	101.72	.000	-1.20
General Information	87.71	105.72	.000	-1.23
Concept Formation	91.56	104.77	.000	-0.97
Analysis-Synthesis	98.15	104.15	.002	-0.44
Visual Auditory Learning	89.21	94.56	.005	-0.39
Retrieval Fluency	87.78	97.43	.000	-0.63
Spatial Relations	98.83	102.78	.004	-0.40
Picture Recognition	103.38	106.33	.049	-0.28
Sound Blending	95.92	107.79	.000	-0.97
Auditory Attention	97.41	104.62	.000	-0.54
Visual Matching	84.54	89.85	.009	-0.37
Decision Speed	97.19	99.49	.264	-0.16
Numbers Reversed	88.97	93.09	.028	-0.31
Memory for Words	85.87	99.86	.000	-1.04

Note. N = 234. ELL = English language learner, MES = monolingual English speaker, WJ-III-NU = Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update.

Table 10

Template of C-LIM Diagnostic Decisions Based on Language Status

		C-LIM Decisions	
		Declining Pattern	No Declining Pattern
Language Status	ELL	TP	FN
	MES	FP	TN

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, MES = monolingual English speaker, TP = true positive, FP = false positive, FN = false negative, TN = true negative.

the most stringent criteria was also visually presented in Figure 3 (p. 20).

No cases followed the declining patterns based on the most stringent (9-cell) interpretation. There were 0 cases for the true positive decisions (i.e., ELLs who followed a

declining pattern), which indicated that a decline in scores in the C-LIM was not evident for students for whom it was expected. In contrast, all of the MES cases were classified under the true negative decisions, indicating that the C-LIM performed as anticipated for these students. Their scores were not expected to follow any of the declining patterns. The frequency counts based on the most stringent criteria (the 9-cell decline) of ELL and MES cases for each C-LIM decision (no declining pattern versus declining pattern) for each of the three patterns of decline are presented in Tables 11-13, respectively.

Table 11

Frequency Count of C-LIM Decision for the Matched Sample Based on Language Status for Influence of Cultural Loading Using the Most Stringent Interpretation

		Most Stringent C-LIM Decisions		
Language Status		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
	ELL (*n*)	0	78	78
	MES (*n*)	0	156	156
	Total	0	234	234

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, MES = monolingual English speaker.

Table 12

Frequency Count of C-LIM Decision for the Matched Sample Based on Language Status for Influence of Linguistic Demand Using the Most Stringent Interpretation

		Most Stringent C-LIM Decisions		
Language Status		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
	ELL (*n*)	0	78	78
	MES (*n*)	0	156	156
	Total	0	234	234

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, MES = monolingual English speaker.

Moderately Stringent C-LIM Interpretation

For the moderately stringent criteria (5-level interpretation), scores from all nine cells in the matrix were averaged according to the five levels presented in Figure 5 (p. 22), and discussed

Table 13

Frequency Count of C-LIM Decision for the Matched Sample Based on Language Status for Combined Influence of Cultural Loading and Linguistic Demand Using the Most Stringent Interpretation

		Most Stringent C-LIM Decisions		
Language Status		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
	ELL (*n*)	0	78	78
	MES (*n*)	0	156	156
	Total	0	234	234

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, MES = monolingual English speaker.

in the Method (pp. 18-19). An equal number of cases from both samples followed the declining pattern based on the singular influence of cultural loading (*n's* = 14 each for ELL and MES) and the singular influence of linguistic demand (*n's* = 7 each for ELL and MES). For the combined influence of cultural loading and linguistic demand, 0 MES and 4 ELL cases followed the declining pattern. Decisions for individual cases based on the moderately stringent criteria for C-LIM interpretation resulted in a low number of true positive decisions (i.e., ELLs followed a declining pattern). Again, a decline in scores in the C-LIM was not evident for students for whom it was expected. However, the C-LIM performed as anticipated for most of the MES cases (no declining pattern), as indicated by a high number of true negative decisions. The moderately stringent C-LIM decisions for both samples based on the three patterns of decline (cultural loading only, linguistic demand only, or combined) are presented in Tables 14-16, respectively.

Least Stringent C-LIM Interpretation

Interpretation of the pattern of decline in scores in the C-LIM according to the least stringent criteria involved a comparison of test scores from only three cells in the C-LIM for each expected pattern of decline. This 3-cell interpretation was presented in Figure 2 (p. 19) and

Table 14

Frequency Count of C-LIM Decision for the Matched Sample Based on Language Status for Influence of Cultural Loading Using the Moderately Stringent Interpretation

		Moderately Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	14	64	78
	MES (*n*)	14	142	156
	Total	28	206	234

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, MES = monolingual English speaker.

Table 15

Frequency Count of C-LIM Decision for the Matched Sample Based on Language Status for Influence of Linguistic Demand Using the Moderately Stringent Interpretation

		Moderately Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	7	71	78
	MES (*n*)	7	149	156
	Total	14	220	234

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, MES = monolingual English speaker.

Table 16

Frequency Count of C-LIM Decision for the Matched Sample Based on Language Status for Combined Influence of Cultural Loading and Linguistic Demand Using the Moderately Stringent Interpretation

		Moderately Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	4	74	78
	MES (*n*)	0	156	156
	Total	4	230	234

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, MES = monolingual English speaker.

is explained in the Method (pp. 19-20).

For the decline based on the singular influence of cultural loading, 21 ELLs followed the pattern (TP), and 103 MESs did not (TN). For linguistic demand, 10 ELLs followed the declining pattern and 144 MESs did not. For the combined influence of cultural loading and linguistic demand, 27 ELLs followed the pattern of decline and 140 MESs did not. The use of the least stringent criteria resulted in a higher number of true positive identifications (i.e., ELLs who followed a declining pattern) than use of the most stringent or moderately stringent criteria. However, contrary to expectations, 65% of the ELL sample did not follow the declining pattern (FN) for combined cultural loading and linguistic demand. Also, of the three types of criteria for C-LIM interpretation, the least stringent criteria resulted in the lowest number of true negative identifications (i.e., MESs who did not follow the pattern) across the three patterns of decline. C-LIM decisions for ELL and MES cases based on the three expected patterns (cultural loading only, linguistic demand only, or combined) using the least stringent interpretation of the C-LIM are presented in Tables 17-19, respectively.

Table 17

Frequency Count of C-LIM Decision for the Matched Sample Based on Language Status for Influence of Cultural Loading Using to the Least Stringent Interpretation

		Least Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	21	57	78
	MES (*n*)	53	103	156
	Total	74	160	234

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, MES = monolingual English speaker.

Decisions for individual cases based on the least stringent criteria, again, resulted in a low number of true positive decisions (i.e., ELLs who followed a declining pattern). Thus, a decline in scores in the C-LIM was not evident for students for whom it was expected. In addition, most of the MES cases fell under the true negative decisions (no decline pattern), indicating that the

Table 18

Frequency Count of C-LIM Decision for the Matched Sample Based on Language Status for Influence of Linguistic Demand Using the Least Stringent Interpretation

		Least Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	10	68	78
	MES (*n*)	12	144	156
	Total	22	212	234

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, MES = monolingual English speaker.

Table 19

Frequency Count of C-LIM Decision for the Matched Sample Based on Language Status for the Combined Influence of Cultural Loading and Linguistic Demand Using the Least Stringent Interpretation

		Least Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	27	51	78
	MES (*n*)	16	140	156
	Total	43	191	234

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, MES = monolingual English speaker.

C-LIM was performing as anticipated for these students.

C-LIM Pattern and Acculturation Level

Follow-up analyses were conducted to investigate the relationship between the presence of a declining pattern in the two language samples and an independent indicator of acculturation, age of entry into the U.S. A summary of frequency counts for this comparison is contained in Table 20. For ease of presentation, cases were grouped into four categories for age of entry. Under each category separate counts were provided each language group. One age of entry category contained all students born in the U.S., which was interpreted as 0 age of entry. The other three age of entry categories were based on participants who were not born in the U.S. and

Table 20

Summary of Frequencies of Language Samples' Patterns of Decline Based on the Three C-LIM Interpretations and Age of Entry into the U.S.

Age of Entry		Born in U.S.: Age 0				Age 0.1 to 3				Age 3.1 to 7				Age 7.1 to 12			
Language Status		ELL		MES		ELL		MES		ELL		MES		ELL		MES	
Pattern of Decline		Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)
Most Stringent	Cultural	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Linguistic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Combined	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Moderately Stringent	Cultural	11	58	14	136	1	3	0	2	0	2	0	0	1	1	0	0
	Linguistic	6	63	7	143	0	4	0	2	1	1	0	0	0	2	0	0
	Combined	3	66	0	150	1	3	0	2	0	2	0	0	0	2	0	0
Least Stringent	Cultural	20	49	51	99	0	4	0	2	0	2	0	0	0	2	0	0
	Linguistic	10	59	11	139	0	4	0	2	0	2	0	0	0	2	0	0
	Combined	26	43	14	136	1	3	1	1	0	2	0	0	0	2	0	0

Note. N = 229. ELL n = 77. MES n = 152. Cases missing Age of Entry = 5. ELL = English language learner, MES = monolingual English speaker, C-LIM = Culture-Language Interpretive Matrix.

organized based on their cognitive developmental level (Piaget, 1962): (a) 0.1 to 3 (sensorimotor to pre-operational); (b) 3.1 to 7 (pre-operational to concrete operations); and (c) 7.1 to 12 (concrete operations to formal operations). Based on the partitioning of the data—language status, decline status, and interpretation criteria, the sample size was not adequate to conduct inferential statistics on the acculturation variable, entry age into the U.S. The sample size was adequate for using inferential statistics only in examining the relationship between age of entry and language status. For this analysis, age of entry was treated as an actual age years. Those born in the U.S. were recorded an age of 0. All other comparisons of the samples for age of entry into the U.S. in regard to the other factors are presented descriptively.

Because of the assumption of homogeneity of variance was violated, an independent samples t-test, with unequal variance, was conducted. The comparison of language status (ELLs and MESs) on age of entry into the U.S. was statistically significant, (t[77.4] = 2.28, $p < .05$, d = 0.45). As expected, ELLs on average entered the U.S. at a later age (M = 0.55, SD = 2.01) than MES students (M = 0.03, SD = 0.27). In essence, ELLs seemed to enter the U.S. just over 6 months later than the MES students.

No descriptive comparisons could be made at any level based on the most stringent interpretation because no cases followed any of the patterns of decline. The remaining descriptive results are shared by age of entry category. Data for age of entry was missing for four cases (ELL = 1, MES = 4).

Ninety percent (n = 69) of the ELLs and 99% (n = 150) of the MESs were born in the U.S. Of the ELLs born in the U.S., 4% to 16% followed a decline pattern under the moderately stringent criteria (cultural = 16%, linguistic = 9%, combined = 4%), and 14% to 38% followed a decline pattern under the least stringent criteria (cultural = 29%, linguistic = 14%, combined =

38%). For the MES group born in the U.S., 0% to 9% followed a decline pattern under the moderately stringent criteria (cultural = 9%, linguistic = 5%, combined = 0%) and 7% to 34% followed a decline pattern under the least stringent criteria (cultural = 34%, linguistic = 7%, combined = 9%).

Five percent (n = 4) of the ELLs and 1% (n = 2) of the MESs entered the U.S. after birth but before age three. Of the ELLs who entered the U.S. prior to age 3 (n = 4), 1 followed two patterns of decline (cultural and combined) under the moderately stringent criteria and 1 followed one pattern of decline (combined) under the least stringent criteria. The remaining three ELLs did not follow any patterns of decline under the moderately stringent or least stringent criteria. Of the MESs who entered the U.S. after birth but before age 3 (n = 2), 1 followed the combined pattern of decline under the least stringent criteria.

Three percent (n = 2) of the ELLs and 0% (n = 0) of the MESs entered the U.S. between ages 3 and 7. Two ELL students entered the U.S. between 3 and 7 years of age, in which 1 of the 2 followed the linguistic pattern of decline under the moderately stringent criteria. Neither of the ELLs followed the patterns of decline under the least stringent criteria. No MESs entered the U.S. between 3 and 7 years of age.

Three percent (n = 2) of the ELLs and 0% (n = 0) of the MESs entered the U.S. between ages 7 and 12. Of the ELLs who entered the U.S. between ages 7 to 12 (n = 2), 1 ELL followed the cultural pattern of decline under the moderately stringent criteria and neither of the ELLs followed any patterns of decline under the least stringent criteria. No MESs entered the U.S. between 7 and 12 years of age.

C-LIM Pattern and Language Proficiency Level

Follow-up analyses were conducted to investigate the relationship between the presence

of a declining pattern in the two language samples and an independent indicator of English language proficiency (i.e., based on students' WIDA scores). Data were grouped and examined based on the language proficiency measured by the WIDA ACCESS for ELLs test. Group mean comparisons, via ANOVA or t tests, were unable to be calculated using the language proficiency data due to inadequate sample sizes. Thus, follow-up analyses are limited to frequency counts.

Across all English language proficiency levels, a majority of ELLs did not follow the patterns of decline in the C-LIM. The following results are presented from the lowest level of English language proficiency (level 1) to the highest (level 6). Both ELLs (n = 2) with the lowest English language proficiency (level 1-entering) did not follow the declining pattern according to the most stringent, moderately stringent, and least stringent criteria in 8 of 9 comparisons. At level 2 (beginning), the majority (n = 3-5) did not follow the pattern of decline in 8 of 9 comparisons. In all 9 comparisons at level 3 (developing), a high frequency of ELLs (n = 15-31) did not follow the declining pattern based on specified criteria. Thirteen ELLs were categorized into level 4 (expanding), and a majority (n = 9-13) did not follow the pattern of decline for all 9 comparisons. At level 5 (bridging), a majority of ELLs (n = 4-5) did not follow expected patterns of decline for all 9 comparisons. No cases were classified at level 6 (reaching), so no comparisons were made at this level. Overall, contrary to expectations, the number of ELLs following the pattern of decline in the C-LIM did not correspond to lower English language proficiency. A summary of this data—ELLs' English language proficiency based on decline pattern and the C-LIM interpretation criteria—is presented in Table 21.

Table 21

Summary of Frequencies of ELLs' Language Proficiency Level Based on Patterns of Decline and the Three Levels of C-LIM Interpretation

WIDA Score		Level 1		Level 2		Level 3		Level 4		Level 5		Level 6	
Pattern of Decline		Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)
Most Stringent	Cultural	0	0	0	0	0	0	0	0	0	0	0	0
	Linguistic	0	0	0	0	0	0	0	0	0	0	0	0
	Combined	0	0	0	0	0	0	0	0	0	0	0	0
Moderately Stringent	Cultural	1	1	3	2	7	24	4	9	1	4	0	0
	Linguistic	0	2	0	5	4	27	2	11	0	5	0	0
	Combined	0	2	2	3	16	15	3	10	0	5	0	0
Least Stringent	Cultural	0	2	2	3	7	24	1	12	0	5	0	0
	Linguistic	0	2	0	5	4	27	4	9	0	5	0	0
	Combined	0	2	1	4	0	31	0	13	0	5	0	0

Note. $N = 56$. WIDA = World-Class Instructional Design and Assessment, ELL = English language learner, MES = monolingual English speaker, C-LIM = Culture-Language Interpretive Matrix.

Diagnostic Utility Statistics and ROC

Diagnostic utility statistics, including sensitivity, specificity, positive predictive values, and negative predictive values, were calculated for the matched sample using the three C-LIM interpretation criteria previously discussed. As before, three separate analyses were conducted under each interpretation criteria: (a) cultural loading, (b) linguistic demand, and (c) the combination of the two. In addition, ROC analyses were conducted to account for the fact that the values for the other diagnostic utility statistics change depending on the cutoff values used for the diagnostic test and the prevalence rate of the target condition (i.e., cultural and linguistic diversity) in the population.

Most Stringent C-LIM Interpretation

The ROC curves for all three patterns of decline (cultural only, linguistic only, cultural and linguistic combined) of the most stringent interpretation were exactly the same and are depicted in Figure 6. The AUC value for all curves was equal to .50. Thus, all three curves followed the diagonal reference line, which was also equal to .50. In essence, using the most stringent C-LIM criteria to distinguish between ELLs and MESs resulted in an accuracy rate of 50%.

Moderately Stringent C-LIM Interpretation

ROC curves when the moderately stringent criteria were used are contained in Figure 7. The AUC values indicated that regardless of the decline pattern examined, use of these criteria resulted in an accuracy rate of 52-55% in differentiating between ELLs and MESs.

Least Stringent C-LIM Interpretation

The ROC curves associated with use of the least stringent criteria for C-LIM interpretation are contained in Figure 8 for the three patterns. AUC values indicated an accuracy

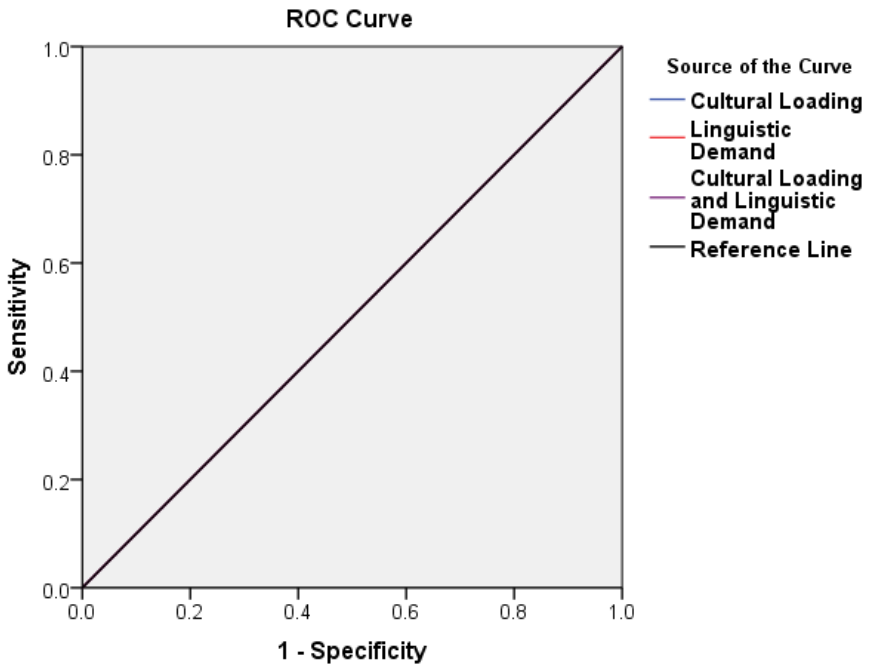

Figure 6. ROC curves for ELLs and MESs based on all three types of decline in WJ-III-NU scores (cultural only, linguistic only, and combined) using the most stringent C-LIM interpretation. ELL = 78; MES = 156.

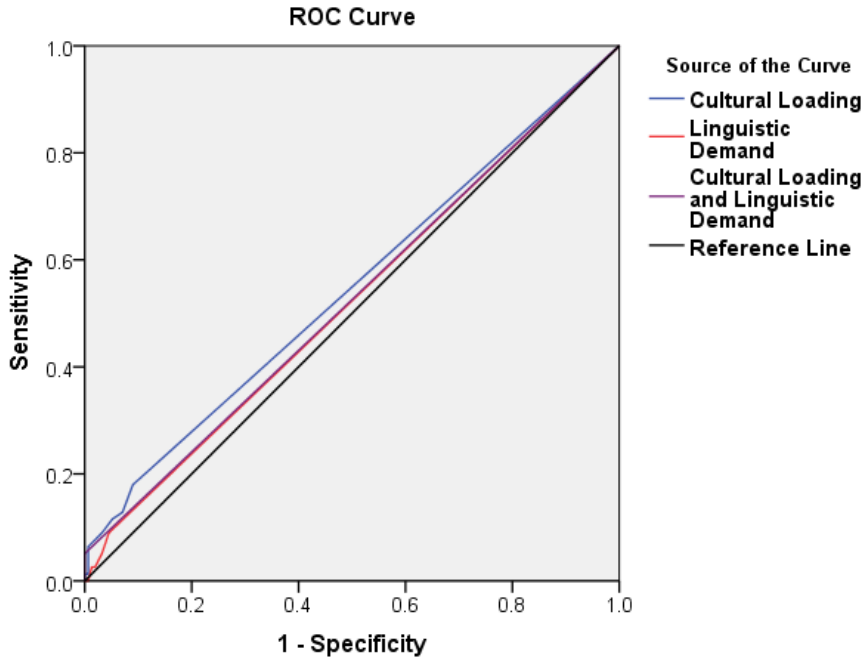

Figure 7. ROC curves for ELLs and MESs based on all three types of decline in WJ-III-NU scores (cultural only, linguistic only, and combined) using the moderately stringent C-LIM interpretation. ELL = 78; MES = 156.

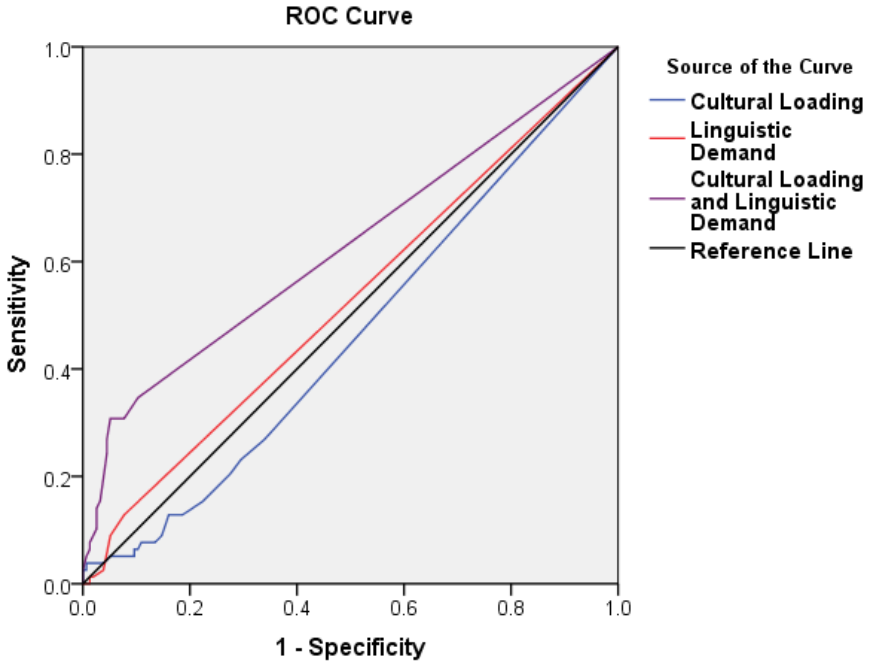

Figure 8. ROC curves for ELLs and MESs based on all three types of decline in WJ-III-NU scores (cultural only, linguistic only, and combined) using the least stringent C-LIM interpretation. ELL = 78; MES = 156.

rate of 46% to 63% in differentiating between ELLs and MESs. The criteria for the three cell decline across the diagonal in the C-LIM resulted in the highest level of accuracy (AUC = .63), relative to other criteria used (AUC = .46 to .55), but remained within the .50 to .70 range, indicating low accuracy of a diagnostic test (Swets, 1988).

Summary

AUC values for all nine types of interpretation (3 criteria x 3 patterns) ranged from 0.46 to 0.63. These finding indicated that all methods of interpretation had a low accuracy rate in distinguishing between ELLs and MESs based on the pattern of decline (Swets, 1988). A summary containing sensitivity, specificity, positive predictive values, negative predictive values, and AUC values for the ROC curves is displayed in Table 22. The set of AUC values presented in Table 22 resulted from ROC analyses using a single continuous variable that was created to represent the TP (ELL with a declining pattern) and FP (MES with a declining pattern)

Table 22

Summary of Diagnostic Utility Statistics for the Matched Sample Based on All Levels of C-LIM Interpretation Criteria

	Pattern of Decline	Sensitivity Value	Specificity Value	PPV Value	NPV Value	AUC
Most Stringent	Cultural Loading	0.00	1.00	undefined	0.67	.50
	Linguistic Demand	0.00	1.00	undefined	0.67	.50
	Combined	0.00	1.00	undefined	0.67	.50
Moderately Stringent	Cultural Loading	0.18	0.91	0.50	0.69	.55
	Linguistic Demand	0.09	0.96	0.50	0.68	.52
	Combined	0.05	1.00	1.00	0.68	.53
Least Stringent	Cultural Loading	0.27	0.66	0.28	0.64	.46
	Linguistic Demand	0.13	0.92	0.45	0.68	.53
	Combined	0.35	0.90	0.63	0.73	.63

Note. N = 234. PPV = positive predictive value, NPV = negative predictive value, AUC = area under curve.

values across all possible ($\geq$ 1) cutoff scores for each type of C-LIM interpretation of the WJ-III-NU scores. Thus, a difference of only 1 point between test scores for each type of C-LIM interpretation was necessary to conclude that a case followed a declining pattern. Sensitivity values were generally low ($\leq$ .35) and specificity values were within the acceptable range ($\geq$ .90), with the exception of the specificity for the decline in cultural loading under least stringent criteria.

Supplemental Binary AUC Calculations

As noted earlier, the first set of AUC values were calculated using a single *continuous* variable for each C-LIM interpretation. This approach limited how or what conclusions could be drawn about the diagnostic utility of the C-LIM. As a result, another set of AUC values was calculated using multiple *dichotomous* variables for each C-LIM interpretation. Thus, the range in diagnostic accuracy of all possible cut scores for the WJ-III-NU data used in this study was attained via these supplemental calculations. Multiple binary AUC values were computed for each cutoff value represented across the range of the continuous variable. The purpose of these calculations was to determine if greater accuracy in C-LIM decisions was achieved through use of a cut score (i.e., minimum difference) other than 1.

No cut scores are offered by Flanagan et al. (2007, 2013) to identify the extent to which scores must differ (e.g., 1 point, 5 points, 10 points) from an expected score (as identified in the C-LIM) in order to count as following a declining pattern. Thus, to determine if a cut point other than one provided greater diagnostic accuracy, each continuous variable was translated into multiple dichotomous variables based on the numerical difference between expected and actual cognitive test scores. To create each dichotomous variable, a binary decision was made with values equal to or above a given value assigned a 1 (i.e., follows declining pattern), and those falling below the given value assigned a 0 (i.e., does not follow the declining pattern).

Findings are presented only for the moderately and least stringent criteria. Based on the most stringent criteria, no cases followed any patterns of decline. Thus, the AUC value would equal .50 no matter what cutoff score was chosen. The AUC values for cutoff scores according to moderately stringent and least stringent criteria are reported in Tables 23 and 24, respectively. The AUC values ranged from .47 to .63 for every possible cut score. No matter what type of interpretation and cut score was used, low diagnostic accuracy was found for the patterns of

Table 23

Percent of Students with Scores Following a Declining Pattern based on Binary AUC Value and Moderately Stringent Criteria

	Cultural Loading Only			Linguistic Demand Only			Cultural Loading & Linguistic Demand		
Cut Score	ELL %	MES %	AUC	ELL %	MES %	AUC	ELL %	MES %	AUC
1	17.9	9.0	.55	9.0	4.5	.52	5.1	0.0	.53
2	12.8	7.1	.53	5.1	3.2	.51	1.3	0.0	.51
3	11.5	5.1	.53	3.8	2.6	.51	0.0	0.0	
4	9.0	3.2	.53	2.6	1.9	.50	0.0	0.0	
5	6.4	0.6	.53	2.6	1.3	.51	0.0	0.0	
6	5.1	0.6	.52	0.0	0.0		0.0	0.0	
7	2.6	0.6	.51	0.0	0.0		0.0	0.0	
8	1.3	0.6	.50	0.0	0.6	.50	0.0	0.0	
9	1.3	0.0	.51	0.0	0.0		0.0	0.0	

Note. N = 234, ELL = 78, MES = 156. ELL = English language learner, MES = monolingual English speaker, AUC = area under curve.

Table 24

Percent of Students with Scores Following a Declining Pattern based on Binary AUC Value and Least Stringent Criteria

	Cultural Loading Only			Linguistic Demand Only			Cultural Loading & Linguistic Demand		
Cut Score	ELL %	MES %	AUC	ELL %	MES %	AUC	ELL %	MES %	AUC
1	26.9	34.0	.47	12.8	7.7	.53	34.6	10.3	.62
2	23.1	29.5	.47	9.0	5.1	.52	30.8	7.7	.62
3	20.5	27.6	.47	2.6	3.8	.49	30.8	5.1	.63
4	15.4	22.4	.47	1.3	1.9	.50	26.9	4.5	.61
5	12.8	18.6	.47	0.0	0.0		24.4	4.5	.60
6	12.8	16.0	.48	1.3	1.3	.50	15.4	3.2	.56
7	9.0	14.7	.47	0.0	0.0		14.1	2.6	.56
8	0.0	0.0		0.0	1.3	.49	10.3	2.6	.54
9	7.7	13.5	.47	0.0	0.0		7.7	1.3	.53
10	7.7	11.5	.48	0.0	0.0		0.0	0.0	
11	7.7	10.9	.48	0.0	0.6	.50	6.4	1.3	.53
12	6.4	10.3	.48	0.0	0.0		5.1	0.6	.52
13	6.4	9.6	.48	0.0	0.0		2.6	0.0	.51
14	5.1	9.6	.48	0.0	0.0		0.0	0.0	
15	5.1	9.0	.48	0.0	0.0		0.0	0.0	
16	5.1	7.1	.49	0.0	0.0		0.0	0.0	
17	5.1	5.1	.50	0.0	0.0		0.0	0.0	
18	0.0	0.0		0.0	0.0		0.0	0.0	
19	0.0	0.0		0.0	0.0		0.0	0.0	
20	3.8	3.8	.50	0.0	0.0		0.0	0.0	
21	0.0	0.0		0.0	0.0		0.0	0.0	
22	0.0	0.0		0.0	0.0		0.0	0.0	
23	3.8	2.6	.51	0.0	0.0		0.0	0.0	

Note. N = 234, ELL = 78, MES = 156. ELL = English language learner, MES = monolingual English speaker, AUC = area under curve.

Table 24 (continued)

Cut Score	Cultural Loading Only			Linguistic Demand Only			Cultural Loading & Linguistic Demand		
	ELL %	MES %	AUC	ELL %	MES %	AUC	ELL %	MES %	AUC
24	3.8	1.9	.51	0.0	0.0		0.0	0.0	
25	3.8	0.6	.52	0.0	0.0		0.0	0.0	
26	0.0	0.0		0.0	0.0		0.0	0.0	
27	0.0	0.0		0.0	0.0		0.0	0.0	
28	0.0	0.0		0.0	0.0		0.0	0.0	
29	2.6	0.6	.51	0.0	0.0		0.0	0.0	
30	0.0	0.0		0.0	0.0		0.0	0.0	
31	0.0	0.0		0.0	0.0		0.0	0.0	
32	0.0	0.0		0.0	0.0		0.0	0.0	
33	0.0	0.0		0.0	0.0		0.0	0.0	
34	0.0	0.0		0.0	0.0		0.0	0.0	
35	0.0	0.0		0.0	0.0		0.0	0.0	
36	2.6	0.0	.51	0.0	0.0		0.0	0.0	
…	0.0	0.0		0.0	0.0		0.0	0.0	
67	1.3	0.0	.51	0.0	0.0		0.0	0.0	

Note. N = 234, ELL = 78, MES = 156. ELL = English language learner, MES = monolingual English speaker, AUC = area under curve.

decline in the C-LIM using the WJ-III-NU scores in delineating ELLs from MESs (Swets, 1988).

Post-Hoc Analyses

A set of post hoc analyses was completed on a subset of students who were pulled from the unmatched sample and who were not diagnosed with SLD (n = 160, ELL = 40, MES = 120). Flanagan et al. (2007) noted, "emerging research suggests that the C-LIM is able to distinguish between culturally and linguistically diverse individuals with and without learning disabilities and thus has some diagnostic utility for practitioners" (p. 182). The purpose of these analyses was to address a potential confounding variable (learning disability status).

Descriptive Statistics

The means and standard deviations for scores on the WJ-III-NU tests for ELL and MES students not identified with SLD are contained in Tables 25 and 26, respectively. On average, the ELL and MES samples did not appear to follow any of the declining patterns (i.e., cultural loading only, linguistic demand only, or combined cultural loading and linguistic demand).

Preliminary Analyses

A series of independent samples t-tests were conducted to determine whether mean differences existed between the two language samples on all of the WJ-III NU tests. For the Visual Auditory Learning test, Levene's Test for Equality of Variances was statistically significant (p = .036), so a t-test with equal variances not assumed was interpreted. Homogeneity of variance was met for all remaining comparisons ($p > .05$). Results of a series of independent-samples t tests indicated that there was a statistically significant difference on 10 out of the 14 tests between the two language samples. The interpretation of the significant findings was the same: The MES students scored higher on the respective test than the ELL students ($p < .02$, d = -1.04 – -0.16). This result was consistent with expectations. A summary of the findings is contained in Table 27.

Table 25

Means and Standard Deviations of WJ-III-NU Subtests for the ELL Sample of Cases not Identified with SLD

		Degree of Linguistic Demand					
		Low		**Moderate**		**High**	
		Test Name	***M (SD)***	**Test Name**	***M (SD)***	**Test Name**	***M (SD)***
Degree of Cultural Loading	**Low**	Spatial Relations	99.4 (9.2)	Numbers Reversed	89.2 (14.2)	Analysis-Synthesis	100.2 (13.3)
				Visual Matching	86.9 (15.7)	Concept Formation	93.9 (13.0)
		Cell Average	**99.4 (9.2)**	**Cell Average**	**88.1 (15.0)**	**Cell Average**	**97.1 (13.2)**
	Moderate	Picture Recognition	104.1 (11.0)	Retrieval Fluency	90.1 (15.2)	Auditory Attention	97.9 (11.5)
				Visual Auditory Learning	91.1 (10.9)	Decision Speed	100.6 (15.4)
						Memory for Words	88.7 (14.3)
						Sound Blending	97.4 (12.0)
		Cell Average	**104.1 (11.0)**	**Cell Average**	**90.6 (13.1)**	**Cell Average**	**96.2 (13.3)**
	High					General Information	92.5 (15.6)
						Verbal Comprehension	90.6 (11.3)
		Cell Average		**Cell Average**		**Cell Average**	**91.6 (13.5)**

Note. n = 40. WJ-III-NU = Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update; ELL =English language learner, SLD = specific learning disability.

Calculations of frequencies for C-LIM decisions. The number of cases resulting from the diagnostic decisions (true positive [TP], false positive [FP], true negative [TN], and false negative [FN]) was calculated based on the C-LIM pattern (declining pattern or no declining pattern) by language status (ELL without SLD or MES without SLD). A template showing the location of frequencies for TP, FP, TN, and FN frequencies is depicted in Table 10 (p. 67). For these analyses, ELLs without SLD were expected to follow the declining patterns and MESs without SLD were expected not to follow the declining patterns. Results of the calculated frequencies are organized based on level of C-LIM interpretation (i.e., most stringent,

Table 26

Means and Standard Deviations of the WJ-III-NU Subtests for the MES Sample of Cases not Identified with SLD

Degree of Cultural Loading	Degree of Linguistic Demand: Low		Moderate		High	
	Test Name	***M (SD)***	**Test Name**	***M (SD)***	**Test Name**	***M (SD)***
Low	Spatial Relations	104.2 (10.6)	Numbers Reversed	95.6 (12.1)	Analysis-Synthesis	105.9 (12.9)
			Visual Matching	89.2 (15.2)	Concept Formation	106.5 (13.4)
	Cell Average	**104.2 (10.6)**	**Cell Average**	**92.4 (13.7)**	**Cell Average**	**106.2 (13.2)**
Moderate	Picture Recognition	105.4 (11.9)	Retrieval Fluency	97.4 (14.5)	Auditory Attention	104.1 (14.0)
			Visual Auditory Learning	95.9 (16.0)	Decision Speed	98.2 (15.6)
					Memory for Words	102. 4(12.9)
					Sound Blending	108.7 (13.4)
	Cell Average	**105.4 (11.9)**	**Cell Average**	**96.7 (15.3)**	**Cell Average**	**103.4 (14.0)**
High					General Information	105.8 (14.8)
					Verbal Comprehension	103.1 (12.4)
	Cell Average		**Cell Average**		**Cell Average**	**104.5 (13.6)**

Note. n = 120. WJ-III-NU = Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update, MES = monolingual English speaker, SLD = specific learning disability.

moderately stringent, least stringent).

Most stringent C-LIM interpretation. For this interpretation, test scores from all nine cells in the C-LIM (refer to pp. 57-58 and Figure 3, p. 20 for further information) were interpreted across (a) cultural loading only, (b) linguistic demand only, and (c) combined cultural loading and linguistic demand. No cases followed the declining patterns based on the most stringent criteria, resulting in 0 true positive identifications (ELLs without SLD with a declining pattern), and 120 true negatives (ELLs without SLD without a declining pattern). Frequency counts of ELL and MES cases for each C-LIM decision (declining vs. no declining pattern)

Table 27

Results of Independent Samples t tests on the WJ-III-NU Scores Based on Language Status for Cases not Identified with SLD

WJ-III-NU Test	ELL (n = 40)	MES (n = 120)	p value	d
Verbal Comprehension	90.60	103.12	.000	-1.04
General Information	92.53	105.82	.000	-0.89
Concept Formation	93.90	106.48	.000	-0.95
Analysis-Synthesis	100.15	105.90	.017	-0.44
Visual Auditory Learning	91.13	95.88	.082	-0.32
Retrieval Fluency	90.10	97.39	.007	-0.50
Spatial Relations	99.35	104.22	.010	-0.48
Picture Recognition	104.10	105.38	.547	-0.11
Sound Blending	97.43	108.72	.000	-0.87
Auditory Attention	97.93	104.06	.013	-0.46
Visual Matching	86.85	89.23	.397	-0.16
Decision Speed	100.55	98.23	.413	0.15
Numbers Reversed	89.18	95.58	.006	-0.51
Memory for Words	88.65	102.36	.000	-1.04

Note. N = 160. ELL = English language learner, MES = monolingual English speaker, SLD = specific learning disability.

based on the 9-cell decline for each of the three patterns of possible decline are contained in Tables 28-30, respectively.

Table 28

Frequency Count of C-LIM Decision Based on Language Status for Influence of Cultural Loading Using the Most Stringent Interpretation for Cases Not Identified with SLD

		Most Stringent C-LIM Decisions		
		Declining Pattern (n)	No Declining Pattern (n)	Total
Language Status	ELL (n)	0	40	40
	MES (n)	0	120	120
	Total	0	160	160

Note. C-LIM = Culture-Language Interpretive Matrix, SLD = specific learning disability, ELL = English language learner, MES = monolingual English speaker.

Table 29

Frequency Count of C-LIM Decision Based on Language Status for Influence of Linguistic Demand Using the Most Stringent Interpretation for Cases Not Identified with SLD

		Most Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	0	40	40
	MES (*n*)	0	120	120
	Total	0	160	160

Note. C-LIM = Culture-Language Interpretive Matrix, SLD = specific learning disability, ELL = English language learner, MES = monolingual English speaker.

Table 30

Frequency Count of C-LIM Decision Based on Language Status for Combined Influence of Cultural Loading and Linguistic Demand Using the Most Stringent Interpretation for Cases Not Identified with SLD

		Most Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	0	40	40
	MES (*n*)	0	120	120
	Total	0	160	160

Note. C-LIM = Culture-Language Interpretive Matrix, SLD = specific learning disability, ELL = English language learner, MES = monolingual English speaker.

Moderately stringent C-LIM interpretation. For the moderately stringent criteria (5-level interpretation), scores from all nine cells in the matrix were averaged into the five levels (refer to Figure 5, p. 22 and pp. 59 for further information). For the singular influence of cultural loading, 6 ELLs without SLD and 16 MESs without SLD followed the pattern. For linguistic demand, 2 ELLs without SLD and 6 MESs without SLD followed the declining pattern. One ELL without SLD and 0 MESs without SLD followed the pattern for the combined influence of cultural loading and linguistic demand. Thus, decisions for individual cases based on the moderately stringent criteria for C-LIM interpretation, again, resulted in a low number of true positive decisions (i.e., ELLs without SLD following a declining pattern). A decline in scores in the C-LIM was not evident for students for whom it was expected. However, the C-LIM

performed as anticipated for most of the MES students without SLD (i.e., no declining pattern), as indicated by a high number of true negative decisions. The C-LIM decisions for the non-SLD, ELL and MES samples based on the three decline patterns are presented in Tables 31-33, respectively.

Table 31

Frequency Count of C-LIM Decision Based on Language Status for Influence of Cultural Loading Using the Moderately Stringent Interpretation for Cases Not Identified with SLD

		Moderately Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	6	34	40
	MES (*n*)	16	104	120
	Total	22	138	160

Note. C-LIM = Culture-Language Interpretive Matrix, SLD = specific learning disability, ELL = English language learner, MES = monolingual English speaker.

Table 32

Frequency Count of C-LIM Decision Based on Language Status for Influence of Linguistic Demand Using the Moderately Stringent Interpretation for Cases Not Identified with SLD

		Moderately Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	2	38	40
	MES (*n*)	6	114	120
	Total	8	152	160

Note. C-LIM = Culture-Language Interpretive Matrix, SLD = specific learning disability, ELL = English language learner, MES = monolingual English speaker.

Table 33

Frequency Count of C-LIM Decision Based on Language Status for Combined Influence of Cultural Loading and Linguistic Demand Using the Moderately Stringent Interpretation for Cases Not Identified with SLD

		Moderately Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	1	39	40
	MES (*n*)	0	120	120
	Total	1	159	160

Note. C-LIM = Culture-Language Interpretive Matrix, SLD = specific learning disability, ELL = English language learner, MES = monolingual English speaker.

Least stringent C-LIM interpretation. Interpretation according to the least stringent criteria involved a comparison of test scores from only three cells in the C-LIM for each expected pattern of decline (i.e., cultural loading only, linguistic demand only, and combined cultural loading and linguistic demand). (Refer to Figure 2 (p. 19) and the explanation of the least stringent criteria in the Method (pp. 59-60) for additional information.) A greater number of non-SLD, ELL and MES cases followed the declining patterns under the least stringent criteria than under the other 2 interpretation criteria. For cultural loading, 12 ELLs without SLD followed the declining pattern (TP), and 73 MESs without SLD did not follow the declining pattern (TN). For linguistic demand, 3 ELLs without SLD followed the declining pattern and 115 MESs without SLD did not follow the declining pattern. For the combined influence of cultural loading and linguistic demand, 10 ELLs without SLD followed the declining pattern and 107 MESs without SLD did not follow the pattern.

Again, the use of the least stringent criteria resulted in a higher number of true positive identifications (i.e., ELLs without SLD following a declining pattern) than use of the two other criteria. However, only 30% of the ELL sample followed the decline for cultural loading, 8% followed the decline for linguistic demand, and 25% followed the decline for the combined influence of cultural loading and linguistic demand. Also, of the three types of criteria for C-LIM interpretation, the least stringent criteria, again, resulted in the lowest number of true negative identifications (i.e., MESs without SLD not following the declining pattern) across all possible decline patterns. C-LIM decisions for non-SLD, ELL and MES cases based on the three decline patterns using the least stringent interpretation of the C-LIM are presented in Tables 34-36, respectively.

C-LIM pattern and acculturation level. Follow-up analyses were conducted to

Table 34

Frequency Count of C-LIM Decision Based on Language Status for Influence of Cultural Loading Using the Least Stringent Interpretation for Cases Not Identified with SLD

		Least Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	12	28	40
	MES (*n*)	47	73	120
	Total	59	101	160

Note. C-LIM = Culture-Language Interpretive Matrix, SLD = specific learning disability, ELL = English language learner, MES = monolingual English speaker.

Table 35

Frequency Count of C-LIM Decision Based on Language Status for Influence of Linguistic Demand According to the Least Stringent Interpretation for Cases Not Identified with SLD

		Least Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	3	37	40
	MES (*n*)	5	115	120
	Total	8	152	160

Note. C-LIM = Culture-Language Interpretive Matrix, SLD = specific learning disability, ELL = English language learner, MES = monolingual English speaker.

Table 36

Frequency Count of C-LIM Decision Based on Language Status for Combined Influence of Cultural Loading and Linguistic Demand According to the Least Stringent Interpretation for Cases Not Identified with SLD

		Least Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
Language Status	ELL (*n*)	10	30	40
	MES (*n*)	13	107	120
	Total	23	137	160

Note. C-LIM = Culture-Language Interpretive Matrix, SLD = specific learning disability, ELL = English language learner, MES = monolingual English speaker.

investigate the relationship between the presence of a declining pattern for the non-SLD sample and an independent indicator of acculturation (U.S. entry age). Because of the small sample size of ELLs, no inferential statistical analyses could be conducted. Thus, only a summary of

frequencies for this comparison is contained in Table 37. For ease of presentation, cases were grouped into four categories for age of entry. Under each category separate counts were provided each language group. One age of entry category contained all students born in the U.S., which was interpreted as 0 age of entry. The other three age of entry categories were based on participants who were not born in the U.S. and organized based on their cognitive developmental level (Piaget, 1962): (a) 0.1 to 3 (sensorimotor to pre-operational); (b) 3.1 to 7 (pre-operational to concrete operations); and (c) 7.1 to 12 (concrete operations to formal operations).

No descriptive comparisons could be made at any level based on the most stringent interpretation because no cases followed any of the patterns of decline. The remaining descriptive results are summarized by age of entry category. Data for age of entry were missing for four cases (MES = 4).

Eighty-five percent (n = 34) of the ELLs and 99% (n = 115) of the MESs were born in the U.S. For the ELL group, 3% to 13% followed a decline pattern under the moderately stringent criteria (cultural = 13%, linguistic = 3%, combined = 3%), and 8% to 30% followed a decline pattern under the least stringent criteria (cultural = 30%, linguistic = 8%, combined = 25%). For the MES group, 0% to 14% followed a decline pattern under the moderately stringent criteria (cultural = 14%, linguistic = 5%, combined = 0%) and 3% to 39% followed a decline pattern under the least stringent criteria (cultural = 39%, linguistic = 3%, combined = 9%). Five percent (n = 2) of the ELLs and 1% (n = 1) of the MESs entered the U.S. after birth but before age three, but none of these participants followed any of the patterns of decline under the moderately stringent or least stringent criteria. Five percent (n = 2) of the ELLs and none of the MESs entered the U.S. between ages 3 and 7. Of the two ELL students who entered the U.S. between these ages, 1 followed the linguistic pattern of decline under the moderately stringent

Table 37

Summary of Frequencies of Age of Entry into the U.S. for each Language Sample without SLD Based on Decline Status for the Three Levels of C-LIM interpretation

Age of Entry		Born in U.S.: Age 0				Age 0.1 to 3				Age 3.1 to 7				Age 7.1 to 12			
Language Status		ELL		MES		ELL		MES		ELL		MES		ELL		MES	
Pattern of Decline		Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)
Most Stringent	Cultural	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Linguistic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Combined	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Moderately Stringent	Cultural	5	29	16	99	0	2	0	1	0	2	0	0	1	1	0	0
	Linguistic	1	33	6	109	0	2	0	1	1	1	0	0	1	1	0	0
	Combined	1	33	0	115	0	2	0	1	0	2	0	0	0	2	0	0
Least Stringent	Cultural	12	22	45	70	0	2	0	1	0	2	0	0	0	2	0	0
	Linguistic	3	31	4	111	0	2	0	1	0	2	0	0	0	2	0	0
	Combined	10	24	11	104	0	2	0	1	0	2	0	0	0	2	0	0

Note. *N* = 156. ELL = 40, MES = 116. Cases missing Age of Entry = 4. ELL = English language learner, MES = monolingual English speaker, C-LIM = Culture-Language Interpretive Matrix.

criteria. Neither of the ELLs followed the patterns of decline under the least stringent criteria. Five percent (n = 2) of the ELLs and 0% (n = 0) of the MESs entered the U.S. between ages 7 and 12. Of the 2 ELLs who entered the U.S. between these ages, 1 followed the cultural and linguistic patterns of decline under the moderately stringent criteria and neither of the ELLs followed any patterns of decline under the least stringent criteria.

C-LIM pattern and language proficiency level. Data were grouped and examined based on English language proficiency as measured by the ACCESS for ELLs test. Insufficient sample size precluded the use of inferential statistics. Thus, a summary of descriptive statistics is presented in Table 38 on English language proficiency for ELLs without SLD based on decline pattern and the C-LIM interpretation criteria. For every level of language proficiency, except for level 6, and across all levels of C-LIM interpretation, except for the most stringent criteria, a majority of ELLs without SLD did not follow the expected pattern of decline (71-100%; n = 20-28). No comparisons could be made at any level based on the most stringent interpretation because no cases met the combined criteria. The same issue existed at level 6 (reaching); no cases met this language proficiency criterion, regardless of the C-LIM interpretation level.

Under the moderately stringent criteria, the largest percent of ELLs without SLD to follow a declining pattern (cultural loading) was approximately 14% across language levels 3 (n = 3) and 4 (n = 1). The largest percent of ELLs without SLD who did not show a declining pattern (combined; 54%; n = 15) were at level 3 (developing) in language proficiency.

In regard to the least stringent criteria, approximately 29% of ELLs followed a declining pattern for cultural loading and were scattered across the language levels (1-5). However, this percent decreased to 7% (n = 2) who had a declining pattern for linguistic demand proficiency;

Table 38

Frequency Distribution of English Language Proficiency Levels Based on the Decline Pattern and the Three Levels of C-LIM Interpretation for ELLs without SLD

WIDA Score		Level 1		Level 2		Level 3		Level 4		Level 5		Level 6	
Pattern of Decline		Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)
Most Stringent	Cultural	0	0	0	0	0	0	0	0	0	0	0	0
	Linguistic	0	0	0	0	0	0	0	0	0	0	0	0
	Combined	0	0	0	0	0	0	0	0	0	0	0	0
Moderately Stringent	Cultural	0	2	0	1	3	12	1	5	0	4	0	0
	Linguistic	0	2	0	1	1	14	0	6	0	4	0	0
	Combined	0	2	0	1	0	15	0	6	0	4	0	0
Least Stringent	Cultural	1	1	1	0	3	12	2	4	1	3	0	0
	Linguistic	0	2	0	1	0	15	2	4	0	4	0	0
	Combined	0	2	0	1	7	8	0	6	0	4	0	0

Note. $N = 28$. ELL = English language learner, SLD = specific learning disability, C-LIM = Culture-Language Interpretive Matrix, WIDA = World-Class Instructional Design and Assessment, Level 1 = entering, Level 2 = beginning, Level 3 = developing, Level 4 = expanding, Level 5 = bridging, Level 6 = reaching.

all were at level 4 proficiency (expanding). However, 25% (n = 7) of the ELLs followed a declining pattern for the combined variables; all were at level 3 proficiency.

Demographic variables were inspected for cases with language proficiency data to determine if the level of language proficiency and decline status was systematically associated with specific demographic characteristics. No consistent pattern could be identified for language proficiency and decline status as a function of age, grade, gender, birth country, eligibility for free or reduced lunch, or age of entry into the U.S.

Diagnostic Utility Statistics

ROC analyses were conducted on the non-SLD sample using the most stringent, moderately stringent, and least stringent criteria previously discussed.

Most stringent C-LIM interpretation. The ROC curves for all three C-LIM levels of the most stringent interpretation were exactly the same, so only one is displayed here (Figure 9). The resulting curve followed the diagonal reference line and the AUC value for all three of the curves was 0.50. Use of the most stringent C-LIM criteria resulted in an accuracy rate of 50% in differentiating between ELLs without SLD and MESs without SLD.

Moderately stringent C-LIM interpretation. ROC curves when the moderately stringent criteria for C-LIM interpretation were used are contained in Figure 10. The AUC values ranged from 0.50 to 0.51, indicating that use of these criteria resulted in an accuracy rate of 50-51% in differentiating between both ELLs and MESs without SLD.

Least stringent C-LIM interpretation. ROC curves using the least stringent criteria for C-LIM interpretation with the non-SLD sample are displayed in Figure 11. AUC values indicated an accuracy rate of 44% to 57% in differentiating between ELLs without SLD and MESs without SLD. The criteria for the three cell decline across the diagonal in the C-LIM

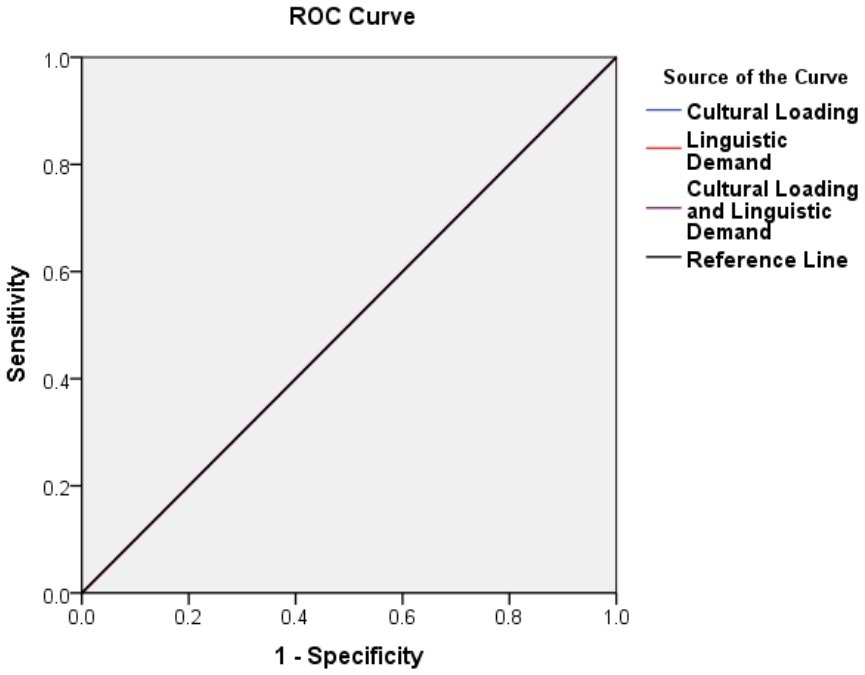

Figure 9. ROC curves for ELLs and MESs in the non-SLD sample based on all three types of decline in WJ-III-NU scores (cultural only, linguistic only, and combined) using the most stringent C-LIM interpretation. ELLs = 40; MESs = 120.

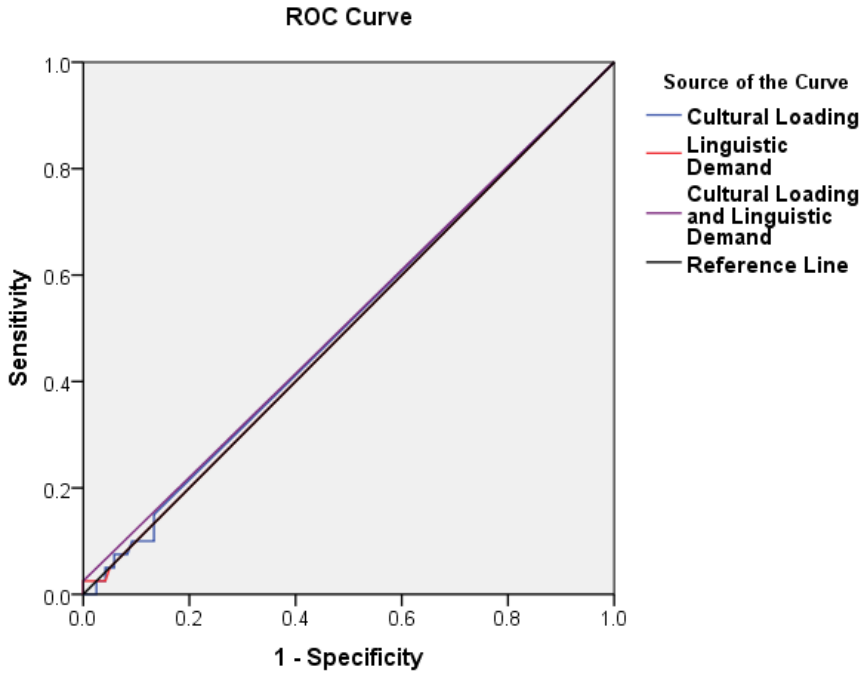

Figure 10. ROC curves for ELLs and MESs in the non-SLD sample based on all three types of decline in WJ-III-NU scores (cultural only, linguistic only, and combined) using the moderately stringent C-LIM interpretation. ELLs = 40; MESs = 120.

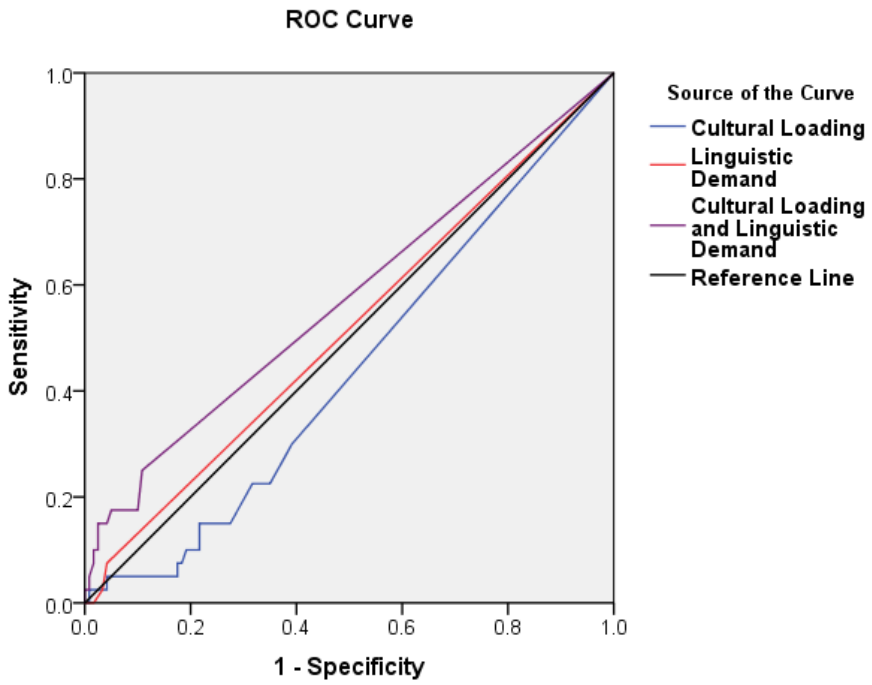

Figure 11. ROC curves for ELLs and MESs in the non-SLD sample based on all three types of decline in WJ-III-NU scores (cultural only, linguistic only, and combined) using the least stringent C-LIM interpretation. ELLs = 40; MESs = 120.

resulted in the highest level of accuracy (AUC = .57), relative to all other criteria (including most stringent and moderately stringent criteria) used (AUC = .44 to .52). Despite being the highest value obtained, the accuracy of the combined influence of cultural loading and linguistic demand based on the least stringent criteria remained within the .50 to .70 range, indicative of low accuracy of a diagnostic test (Swets, 1988).

Summary. AUC values for the nine types of interpretation (3 criteria x 3 patterns) ranged from 0.44 to 0.57, indicating that the application of these C-LIM interpretive methods to the WJ-III-NU scores resulted in low accuracy in distinguishing between ELLs and MESs without SLD (Swets, 1988). A summary table containing sensitivity, specificity, positive predictive values, negative predictive values, and AUC values for the ROC curves depicted in Figures 9-11 is displayed in Table 39. In general, sensitivity values were low (≤ .30) and specificity values were within the acceptable range (≥ .87), with the exception of the specificity

Table 39

Summary of Diagnostic Utility Statistics for the Non-SLD Sample on all levels of C-LIM Interpretation Criteria

	Pattern of Decline	Sensitivity Value	Specificity Value	PPV Value	NPV Value	AUC
Most Stringent	Cultural Loading	0.00	1.00	undefined	0.75	.50
	Linguistic Demand	0.00	1.00	undefined	0.75	.50
	Combined	0.00	1.00	undefined	0.75	.50
Moderately Stringent	Cultural Loading	0.15	0.87	0.27	0.75	.51
	Linguistic Demand	0.05	0.95	0.25	0.75	.50
	Combined	0.03	1.00	1.00	0.75	.51
Least Stringent	Cultural Loading	0.30	0.61	0.20	0.72	.44
	Linguistic Demand	0.08	0.96	0.38	0.76	.52
	Combined	0.25	0.89	0.43	0.78	.57

Note. N = 160. SLD = specific learning disability, PPV = positive predictive value, NPV = negative predictive value, AUC = area under curve.

value for the decline in cultural loading under the least stringent criteria.

Supplemental Binary AUC Calculations

The set of AUC values in Table 39 resulted from ROC analyses using a single *continuous* variable for C-LIM interpretation. For the supplementary calculations, multiple binary AUC values were obtained through separate computations for each value represented in the continuous variable. The purpose of these calculations was to determine if greater accuracy in C-LIM decisions was achieved through use of a cut score (i.e., minimum difference) other than 1. Thus,

for each degree of difference (i.e., cutoff score), a binary decision was made with values equal to or above a given value assigned a 1 (i.e., follows declining pattern), and those falling below the given value assigned a 0 (i.e., does not follow the declining pattern). The range in diagnostic accuracy of all possible cut scores for the non-SLD sample was attained via these supplemental calculations.

Zero cases followed any patterns of decline based on the most stringent interpretation; thus, the AUC value would equal .50 no matter what cutoff score was chosen. AUC values for cutoff scores according to the moderately stringent and least stringent criteria for the non-SLD sample are reported in Tables 40 and 41, respectively. Again, the AUC values ranged from .44 to .57 for every possible cut score. Interpretation of patterns of decline in the C-LIM demonstrated low diagnostic accuracy in delineating ELLs without SLD from MESs without SLD no matter what type of interpretation or cut score was used (Swets, 1988). Contained in the Appendix is an additional set of post-hoc analyses that were completed on a sample of ELLs only, who were separated into groups based on the presence or absence of SLD.

Table 40

Percent of Non-SLD Students with a Decline Pattern based on Binary AUC Value and Moderately Stringent Criteria

	Cultural Loading Only			Linguistic Demand Only			Cultural Loading & Linguistic Demand		
Cut Score	ELL %	MES %	AUC	ELL %	MES %	AUC	ELL %	MES %	AUC
1	15.0	13.3	.51	5.0	5.0	.50	0.0	0.0	
2	10.0	11.7	.49	2.5	4.2	.49	2.5	0.0	.51
3	7.5	8.3	.50	2.5	2.5	.50	0.0	0.0	
4	5.0	4.2	.50	2.5	1.7	.50	0.0	0.0	
5	2.5	2.5	.50	2.5	0.8	.51	0.0	0.0	
6	0.0	2.5	.49	0.0	0.0		0.0	0.0	
7	0.0	1.7	.49	0.0	0.0		0.0	0.0	
8	0.0	0.8	.50	0.0	0.0		0.0	0.0	

Note. N = 160, ELL = 40, MES = 120. AUC = area under curve, SLD = specific learning disability, ELL = English language learner, MES = monolingual English speaker.

Table 41

Percent of Non-SLD Students with a Decline Pattern based on Binary AUC Value and Least Stringent Criteria

	Cultural Loading Only			Linguistic Demand Only			Cultural Loading & Linguistic Demand		
Cut Score	ELL %	MES %	AUC	ELL %	MES %	AUC	ELL %	MES %	AUC
1	30.0	39.2	.45	7.5	4.2	.52	25.0	10.8	.57
2	22.5	35.0	.44	2.5	3.3	.50	17.5	10.0	.54
3	22.5	31.7	.45	0.0	1.7	.49	17.5	5.0	.56
4	15.0	27.5	.44	0.0	0.0		15.0	4.2	.55
5	15.0	24.2	.45	0.0	0.0		15.0	2.5	.56
6	15.0	21.7	.47	0.0	0.0		10.0	2.5	.54
7	12.5	21.7	.45	0.0	0.0		10.0	1.7	.54
8	0.0	0.0		0.0	0.0		7.5	1.7	.53
9	10.0	21.7	.44	0.0	0.0		5.0	0.8	.52
10	0.0	0.0		0.0	0.0		0.0	0.0	
11	10.0	19.2	.45	0.0	0.0		2.5	0.8	.51
12	7.5	18.3	.45	0.0	0.0		0.0	0.0	
13	7.5	17.5	.45	0.0	0.0		2.5	0.0	.51
14	5.0	17.5	.44	0.0	0.0		0.0	0.0	
15	5.0	15.8	.45	0.0	0.0		0.0	0.0	
16	5.0	10.8	.47	0.0	0.0		0.0	0.0	
17	5.0	8.3	.48	0.0	0.0		0.0	0.0	
18	0.0	0.0		0.0	0.0		0.0	0.0	
19	0.0	0.0		0.0	0.0		0.0	0.0	
20	5.0	7.5	.49	0.0	0.0		0.0	0.0	
21	0.0	0.0		0.0	0.0		0.0	0.0	

Note. N = 160, ELL = 40, MES = 120. AUC = area under curve, SLD = specific learning disability, ELL = English language learner, MES = monolingual English speaker.

Table 41 (continued)

	Cultural Loading Only			Linguistic Demand Only			Cultural Loading & Linguistic Demand		
Cut Score	ELL %	MES %	AUC	ELL %	MES %	AUC	ELL %	MES %	AUC
22	0.0	0.0		0.0	0.0		0.0	0.0	
23	5.0	6.7	.49	0.0	0.0		0.0	0.0	
24	5.0	5.8	.50	0.0	0.0		0.0	0.0	
25	5.0	4.2	.50	0.0	0.0		0.0	0.0	
26	0.0	0.0		0.0	0.0		0.0	0.0	
27	2.5	4.2	.49	0.0	0.0		0.0	0.0	
28	2.5	3.3	.50	0.0	0.0		0.0	0.0	
29	2.5	2.5	.50	0.0	0.0		0.0	0.0	
30	0.0	0.0		0.0	0.0		0.0	0.0	
31	0.0	0.0		0.0	0.0		0.0	0.0	
32	0.0	0.0		0.0	0.0		0.0	0.0	
33	0.0	0.0		0.0	0.0		0.0	0.0	
34	2.5	1.7	.50	0.0	0.0		0.0	0.0	
35	0.0	0.0		0.0	0.0		0.0	0.0	
36	2.5	0.8	.51	0.0	0.0		0.0	0.0	
…	0.0	0.0		0.0	0.0		0.0	0.0	
45	0.0	0.8	.50	0.0	0.0		0.0	0.0	

Note. N = 160, ELL = 40, MES = 120. ELL = English language learner, MES = monolingual English speaker, AUC = area under curve.

DISCUSSION

The purpose of the study was to examine the utility of the C-LIM as a measure of the validity of cognitive test scores for English language learners. Flanagan et al. (2013) asserted that ELLs would generally demonstrate a pattern of decline in their cognitive scores when entered into the C-LIM due to differences in language proficiency and acculturation from the mainstream, English-only speaking samples on which cognitive tests are commonly normed. This theory was tested in several ways. The expected decline in scores in the C-LIM was examined through identifying the patterns of cognitive performance for ELLs and then in comparing their patterns to that of monolingual, English-speaking students. Then, language proficiency and acculturation levels of the ELLs were inspected to determine the relationship between these variables and the presence of a declining C-LIM pattern. The findings did not support the C-LIM pattern for ELLs. Thus, the discussion begins with a summary of what was examined within the C-LIM model, followed by an explanation of results with respect to extant literature. The discussion ends with a description of the limitations of the study, recommendations for future research, and implications for practice.

Expected Patterns of Performance

The study was designed to investigate several aspects of the C-LIM, including different levels of interpretation, assumptions regarding language proficiency and acculturation, and limitations of prior research. In regard to interpretation, three types of criteria (varying in stringency) across the three levels of C-LIM interpretation (singular effects of cultural loading and linguistic demand, and the combined effect) resulted in a total of nine ways in which the C-LIM was investigated. Based on the assertions of Flanagan and colleagues (2007, 2013), it was expected that the cognitive performance of ELL students would generally follow three possible

declining patterns, which would fall within one of three types of interpretation criteria, while the performance of MES students would not follow any of the patterns.

In regard to acculturation and English language proficiency, Flanagan et al. (2007, 2013) assert that the declining C-LIM pattern should only be evident in cases where, other than true cognitive ability, the student's level of acculturation and English language proficiency are the primary influences on his or her performance. This assertion was examined through inspection of data related to acculturation and English language proficiency with regard to whether cases followed or did not follow the expected patterns of performance. Independent examinations of cultural loading and linguistic demand are necessary to ensure that the constructs are viable for interpretation in the C-LIM as operationalized.

The study also addressed limitations of prior research. Research conducted after the C-LIM was developed (Aziz, 2009; Brown, 2008; Cormier, 2012; Dhaniram-Beharry, 2008; Kranzler et al., 2010, Lella Souravlis, 2010; Nieves-Brull, 2006; Templeton, 2012; Tychanska, 2009; Verderosa, 2007) has used primarily nomothetic (study of groups; Allport, 1937) methods to assess the validity of the C-LIM. In contrast, the primary intent of this study was to evaluate the C-LIM the way it is used in practice, on an individual, or idiographic (study of individuals; Allport, 1937) basis. Findings resulting from the idiographic approach are more important in that this approach is consistent with the way the C-LIM is used in practice.

The C-LIM and Diagnostic Utility

Diagnostic utility statistics are used to investigate whether diagnostic tests work as expected. Specifically, these statistics are used to answer the following question: How accurately does a test identify a condition in those who have it versus those who do not? Most importantly, diagnostic utility statistics allow for examination of how the C-LIM works on an

individual basis, and thus are an appropriate approach for determining its adequacy for use in practice.

In this study, diagnostic utility statistics were used to determine the accuracy of the C-LIM in identifying cultural and linguistic diversity in ELLs. ROC curve analysis was conducted because it allows for determinations of accuracy across all possible cut scores of a test, and is not influenced by prevalence rate of a condition in the population (McFall & Treat, 1999). The area under the curve (AUC) value for each ROC curve was examined as a measure of the rate at which the C-LIM distinguished between ELLs and monolingual, English-speakers based on the presence of a declining pattern.

ELL to MES Comparison

At an idiographic level, C-LIM interpretation based on all levels of criteria resulted in no or a low number of ELL cases following the declining pattern and a high number of MES cases not following the pattern. Under the most stringent C-LIM interpretation, none of the cases, regardless of language status (ELL or MES) followed the patterns of decline for any singular or combined influence. The lack of pattern for the MES students is expected based on the premise of C-LIM, but the findings are not in alignment with the C-LIM for the ELL students. Although the moderately stringent criteria resulted in an increase of ELL cases following the declining pattern (TP), the maximum percent was less than 20%. As sensitivity (TP) increased, specificity (TN) decreased (MES with no declining pattern), with the percentage of FP (MES cases with a declining pattern) in the same range as TPs, while the percentage of true negatives decreased. Thus, support for C-LIM under moderate interpretation criteria is mixed; the value of TN for predicting MES's is acceptable for most criteria, but TP is inadequately low in predicting the pattern for ELLs (Matthey & Petrovski, 2002). When the least stringent C-LIM interpretation

criteria were followed, 13% to 35% of ELLs and 8% to 34% of MESs followed the three patterns of decline. Using a similar analytical approach, Styck (2012) found that 11% of ELL students followed the declining pattern. Taken together, these findings provide little support for the C-LIM. Results of diagnostic utility statistics indicated that, for all nine possible C-LIM interpretations, a large percentage of MES students did not follow the declining pattern. Thus, C-LIM interpretation generally resulted in correct decisions for cases that should not have followed a declining pattern. However, C-LIM interpretation also resulted in a high number of ELLs who were incorrectly identified because their scores did not follow the declining patterns.

As a whole, these findings suggest that the C-LIM does not function as expected on an individual level to differentiate between ELLs and MESs. These findings are consistent with prior research (Styck, 2012) conducted using diagnostic utility statistics and ROC curve analysis with another cognitive test. Styck (2012) reported that the C-LIM demonstrated low sensitivity, high specificity, and low diagnostic accuracy in distinguishing between ELL students and a MES sample (the WISC-IV standardization sample). Although depressions in cognitive scores for ELLs as compared to MESs have been found on average in previous research (Cummins, 1984; Goddard, 1917; Jensen, 1974, 1976; Mercer, 1979; Nieves-Brull, 2006; Sánchez, 1934; Valdés & Figueroa, 1994; Verderosa, 2007), these findings are not sufficient to conclude that the same depression in cognitive scores will be evident when the C-LIM is used for individual ELLs. The fact that the C-LIM did not function on an individual level in this study the same as it has for group mean comparisons is not entirely surprising, given that mean score differences are not adequate evidence to support the presence of individual differences (Watkins, 2003). Prior researchers (e.g. Brown, 2008; Cormier, 2012; Nieves-Brull, 2006) who have used nomethetic

analyses also found that the C-LIM has not functioned as expected, and have suggested revisions to the matrix based on their findings.

Despite the lack of support for the C-LIM based on ELL to MES comparisons, characteristics of the sample used in this study may have influenced the lack of support for the C-LIM interpretation. Thus, in an attempt to address a few potential confounds, and potentially increase confidence in the findings from the primary analyses, additional analyses were conducted.

Influences of Acculturation and Language Proficiency

Acculturation in this study was operationally defined as the age of entry into the U.S. No other acculturation variables (e.g., native language proficiency, bilingual proficiency, national origin, percent of student population speaking same first language) were accessible in the school's archival database. Given the available acculturation data, the sample size of students who entered the U.S. after birth was too small for inferential analyses. As a result, no clear descriptive pattern could be discerned regarding ELLs' scores on the WJ-III-NU to support that ELLs who entered the U.S. at a later age tended to follow a declining pattern in the C-LIM.

The ACCESS for ELLs language proficiency test was used to define the levels of English language proficiency for ELLs in this study. The hypothesis that ELLs who exhibited a declining pattern on the C-LIM would also demonstrate lower English language proficiency could not be supported. No clear descriptive pattern could be discerned between ELLs' WJ-III-NU scores and English language proficiency. At all levels of English language proficiency, the cognitive scores for a majority of the ELLs did not follow any of the declining patterns.

One possible explanation for the lack support of the C-LIM in this study is that the ELL students were no different than the MES students in their level of acculturation or language

proficiency. No independent or direct comparison could be made between these two language groups on acculturation or language skills, because the MES students are presumed to be culturally adept and proficient in English, and thus are not administered acculturation or language proficiency measures.

Another explanation for the acculturation findings could be due to the measurement of this variable. Age of entry into the U.S. was used as the proxy for acculturation. Use of a single indicator of acculturation is problematic because other factors, such as number of years spent in the U.S., number of years receiving language services in an ELL program, English language proficiency, and national origin contribute to level of acculturation (Collier, 2001). Despite these potential weaknesses, this study is one of two known studies (i.e., Verderosa, 2007) that have examined the relationship between language proficiency and the C-LIM pattern. It may also be one of the few studies, if not the first, to measure acculturation independently of the C-LIM matrix.

Post-Hoc Analyses: ELL to MES Comparison - Non-SLD Only

According to Flanagan et al. (2007, 2013), if a student has a learning disability, it is assumed that his or her cognitive strengths and weaknesses would result in a distribution of scores that would not follow a declining C-LIM pattern. Because the sample consisted of referred MESs and Spanish-speaking ELLs, post hoc analyses were conducted to rule out the potential effects of learning disability on the students' cognitive performance on the WJ-III-NU.

For students without learning disabilities, C-LIM interpretation based on all levels of criteria resulted in a low number of ELL cases following the declining pattern and a high number of MES cases not following the pattern. Because students with SLD were removed from the sample, the expectation was that a larger percentage of the ELL group may follow the declining patterns examined. However, the percentage of ELL students following declining patterns was

lower in the non-SLD sample. Also, contrary to expectations, the percentage of MES students that followed the declining pattern generally increased. In previous research (Kranzler et al., 2010) on ELLs without identified disabilities, 37% of ELL students followed the declining pattern. This percent is higher than the percent found in this study (25%) under similar criteria.

The C-LIM demonstrated low diagnostic accuracy, via ROC curve analysis and AUC values, in differentiating between ELL and MES students without SLD no matter what criteria were used. Despite controlling for a potential confound (i.e., presence of a learning disability), the results for the non-SLD sample were even poorer than the results for the sample in which students with SLD were included.

Finally, it is notable that the conception of the C-LIM was based on studies that generally used non-referred samples (Cummins, 1984; Goddard, 1917; Jensen, 1974, 1976; Mercer, 1979; Sanchez, 1934; Valdéz & Figueroa, 1994; Vukovich & Figueroa, 1982), but the C-LIM is recommended for use during special education evaluations of referred, English language learners (Flanagan et al., 2007, 2013). Thus, the findings of this study are potentially representative of the type of results school psychologists may find if the C-LIM is used and diagnostic accuracy statistics are completed on a sample of referred ELL and MES students within their district.

Limitations of Current Study

Characteristics of the sample used in this study contributed to several limitations. The sample was geographically restricted, and thus not nationally representative. Also, the use of a convenience sample limited the type and amount of data that could be obtained. For example, detailed information regarding level of acculturation was unavailable. Using the length of time in the U.S. is an imprecise measure of acculturation and limits the conclusions that can be drawn regarding the effect of acculturation on cognitive performance. Furthermore, just as many ELLS like MESs were born in the U.S. Thus, the overall lack of variation in acculturation level in the

sample did not allow for the best comparisons between groups on this construct. Data on language proficiency were also restricted because only the ELLs in the sample were examined for English language proficiency. English speakers do not usually take an examination of proficiency in their native language. In addition, the acculturation and English language proficiency levels demonstrated by referred students may be different than those demonstrated by non-referred students. Use of a primarily Hispanic, Spanish-speaking sample also limits the generalizability of findings to culturally and linguistically similar individuals. Although data were gathered from students with different cultural and linguistic backgrounds, the sample size for each group was inadequate for further analysis. Finally, because ELL students were referred as a result of educational difficulties, they may not be representative of the ELLs for whom declining C-LIM patterns are expected (i.e., non-disabled ELLs). When an attempt was made to account for this potential confound by removing students with learning disabilities from the sample, the diagnostic accuracy of C-LIM decisions was practically the same. However, a more direct comparison is needed between an adequate sample size of ELLs and non-ELLS across the spectrum of learning statuses in order to draw conclusions about the actual relationship between the C-LIM and learning abilities.

Another limitation of the current study was that the sample was obtained from a school district in which the C-LIM is used during evaluations of ELLs. As a result, any evaluations that used the C-LIM to determine the validity of test results in making decisions about the presence of disability may have skewed the results of this study by creating a self-fulfilling prophecy. Essentially, it would not be surprising if the performance of ELLs without SLD followed a declining pattern and ELLs with SLD did not follow a declining pattern, because the C-LIM contributed to those decisions. However, findings that separated students based on SLD status

still demonstrated low diagnostic accuracy. Again, a more direct comparison is needed between ELL cases in which the C-LIM was used to make decisions about the viability of the cognitive scores and those cases in which the C-LIM was not used.

Finally, although an attempt was made to proportionally match ELLs and MESs based on demographic variables, the groups were only matched based on age and gender to avoid losing a significant number of cases. As a result, group differences on characteristics such as race/ethnicity, socioeconomic status (SES), and GIA could have influenced the findings.

Recommendations for Future Research

There are several avenues for future research on factors influencing the validity of cognitive test performance for ELLs and the methods used to address such factors. For example, English language proficiency and acculturation are areas that can influence ELL students performance on cognitive measures, but the specific effects of these variables on cognitive test performance have not been adequately researched. One approach toward research in this area could involve separation of participants into different groups based on levels of direct (e.g., language proficiency test) and indirect (e.g., self- or parent-report acculturation survey) measures of acculturation and language proficiency. Rather than attempting to confirm that ELL students' performance follows any of the declining C-LIM patterns, research in this area should be exploratory in order to determine what tests, if any, demonstrate a clinically significant depression in performance between groups (without entering scores into the C-LIM). Then attempts should be made to replicate findings with additional samples and research conducted by third parties. A consistent pattern in performance for ELLs, on an individual level, on specific cognitive tests (as compared to test norms) that meets requirements for diagnostic accuracy of a high-stakes assessment would be necessary to support the application of such knowledge in practice.

There is a paucity of measures that have been translated, standardized, and normed for use with individuals who speak languages other than English or Spanish. Additional research needs to be conducted to support the translation and cultural adaptation of cognitive, academic, behavioral and social-emotional measures used during educational evaluations. Future research should also focus on the development and effective use of multicultural competencies by school psychologists (Belar, 2009; Jones, Sander, & Booker, 2013; Lopez & Bursztyn, 2013; Newell et al., 2010). Such training is essential in order for practitioners to not only be aware of areas that should be assessed during evaluations of ELLs, but also how to approach ELL assessment in a systematic and comprehensive manner given the potential personal differences (language, culture, etc.) between the examiner and examinee, and limitations of currently available assessment methods.

Implications for Practice

Flanagan et al. (2007) asserted, "use of XBA along with the C-LTC and C-LIM provides a systematic and defensible method for greatly reducing the discriminatory aspects inherent in the use of cognitive ability tests with diverse individuals" (p. 167). The results of this study do not support that conclusion. Instead, the C-LIM, used as a tool for interpreting the WJ-NU-III performance of English language learners and MES students, demonstrated a level of diagnostic accuracy much lower than what is considered acceptable for clinical decisions made on an individual basis (Swets, 1988). Despite the limitations of this study, the findings clearly raise a red flag regarding the current use of the C-LIM by school psychologists in their daily practice.

To date three independent studies (Kranzler, et al., 2010; Styck, 2012), including this one, have been conducted that have raised concerns about the viability of the C-LIM and its use in clinical practice. Each study has approached testing the C-LIM in different ways. Kranzler and colleagues examined frequencies of three C-LIM patterns in WJ-III-NU scores for a sample

of non-referred ELLs. Styck examined the diagnostic utility of the WISC-IV in distinguishing between referred ELLs grouped based on different SLD criteria and the WISC-IV normative sample. Finally, in this study, the diagnostic utility of the C-LIM with the WJ-III-NU was examined in comparisons between Spanish-speaking ELLs and monolingual English-speaking students in a referred sample. The influences of acculturation, language proficiency, and disability status on C-LIM patterns were also investigated. Utilization of cognitive assessments commonly administered by school psychologists and variation in the way the C-LIM was studied only strengthens the cumulative evidence about the current viability and usage of the C-LIM in practice. At this time, the C-LIM is not recommended for use in evaluating the validity of cognitive test performance for ELLs.

Instead, practitioners are encouraged to follow educational laws (Americans with Disabilities Act of 1990; Education for All Handicapped Children Act of 1975), and professional codes of ethics (*Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations*, American Psychological Association, 1990; *Principles for Professional Ethics*, National Association of School Psychologists, 2010; Rogers et al., 1999; *Standards for Educational and Psychological Testing*, American Educational Research Association, APA, & National Council on Measurement in Education, 1999) when assessing all students, including ELLs. The use of current "Best Practices in Nondiscriminatory Assessment" (Ortiz, 2008) is also recommended.

When assessing ELLs, as with all comprehensive evaluations for educational disabilities, multiple factors need to be considered that could be influencing the students' educational performance. For ELLs, language proficiency (in both languages), exposure to language at home and in school, culture of origin, generational status, familial customs, and level of acculturation

are some of the factors that should be investigated (Flanagan et al., 2007; Ortiz, 2008; Rogers-Adkinson, Ochoa, & Delgado, 2003). Although the C-LIM was conceptualized to address some of these factors in psychoeducational evaluations, it does not operate as expected and thus should not be used as a method to account for such factors. As a result, school psychologists are tasked with making sure level of acculturation and language proficiency are considered as potential explanations for ELLs' educational performance via examination of independent indicators of these factors. To provide an appropriate, comprehensive assessment of an ELL student's cognitive abilities, acculturation and English language proficiency must be considered a priori (Ortiz, 2008). Also, the administration of a nonverbal assessment of cognitive ability is recommended along with assessment in other cognitive areas (Lakin, 2012). If other measures are administered, it is important to monitor the ELL's performance on measures with high verbal demand (e.g., General Information, Verbal Comprehension), as ELLs have consistently demonstrated significantly lower performance on verbal tests (Bialystok & Craik, 2010; Nieves-Brull, 2006; Verderosa, 2007; Styck, 2012). Monitoring performance includes informally exploring the ELLs' comprehension of tests with high expectations for cultural knowledge and English language proficiency once the standardized administration has been completed (Ochoa, 2008). If possible, tests that have been developed and normed in the student's first language should also be administered (Sanchez et al., 2013).

To adequately address all of the educational needs of ELLs, school psychologists must be aware of and knowledgeable about the potential influences on ELLs' educational performance (e.g., acculturation, language proficiency, cultural differences regarding language maintenance, cognitive strengths and weaknesses demonstrated by bilingual individuals; Jones, Sander & Booker, 2013; Mindt, 2008; Ortiz, 2008). For school psychologists to gain such awareness,

knowledge, and skills in working with culturally and linguistically diverse populations, school psychology programs must provide training in these areas (Newell, et al., 2010). Furthermore, once trained, it is essential that psychologists disseminate their knowledge to other educators so that ELLs can be better served within general education settings and referrals for cognitive testing as part of special education evaluations are truly a last resort.

Conclusion

Flanagan et al. (2007) created the C-LIM to assist school psychologists to address in a systematic way the potential impact of acculturation and English language proficiency levels on the standardized cognitive test scores of ELLs. However, research to support the use of the matrix as an assessment method for diagnostic purposes is weak, as evidenced in this study, or mixed at best (Nieves-Brull, 2006; Verderosa, 2007). Such research needs to be conducted with rigor using methods that correspond to the manner in which the method is used in practice. The findings of this study and other recent research (Kranzler, et al., 2010, Styck, 2012) do not support the use of the C-LIM. Given that these few studies are the only investigations on the C-LIM as is it used in practice, it is questionable whether C-LIM interpretation could be a viable method for assessing the validity of ELLs' cognitive test scores.

More work needs to be done to develop and validate evidence-based nondiscriminatory methods for cognitive assessment of ELLs. Recommendations from Flanagan et al. (2007) and Ortiz (2008), in combination with laws and professional guidelines, delineate factors important to consider in ELL evaluations. The importance of providing appropriate assessments for ELLs is emphasized through the high-stakes impact of cognitive assessments used during special education evaluations. Inappropriate use of cognitive assessments can adversely affect ELLs' educational placement and ultimate educational success. Thus, diligent efforts from researchers,

school psychology programs, and practitioners are critical to ensure that assessment practices for culturally and linguistically diverse learners continue to improve.

REFERENCES

Alfonso, V. C., Flanagan, D. P., & Radwan, S. (2005). The impact of the Cattell-Horn-Carroll theory on test development and interpretation of cognitive and academic abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 185-202). New York, NY: Guilford Press.

Allport, G. W. (1937). *Personality: A psychological interpretation.* Oxford, England: Holt.

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

American Psychological Association. (1990). *Guidelines for providers of psychological services to ethnic, linguistic, and culturally diverse populations.* Washington, DC: Author. Retrieved from http://www.apa.org/pi/oema/resources/policy/provider-guidelines.aspx

Americans With Disabilities Act of 1990, 42 U.S. C. A. § 12101 *et seq.* (West 1993).

Artiles, A. J., Rueda, R., Salazar, J. J., & Higareda, I. (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. *Exceptional Children, 71,* 283-300. Retrieved from .psu.edu/docview/201200908/fulltext PDF?accountid=13158

Artiles, A. J., & Trent, S. C. (1994). Overrepresentation of minority students in special education: A continuing debate. *Journal of Special Education, 27,* 410-437. doi: 10.1177/002246699402700404

Aud, S., Hussar, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., & Tahan, K. (2011). *The Condition of Education 2011* (NCES 2011-033). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Retrieved from http://nces.ed.gov/pubs2011/2011033.pdf

Aziz, N. (2009). *Patterns of cognitive performance for culturally and linguistically diverse individuals with global cognitive impairment.* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3441046)

Bainter, T. R., & Tollefson, N. (2003). Intellectual assessment of language minority students: What do school psychologists believe are acceptable practices? *Psychology in the Schools, 40,* 599-603. doi: 10.1002/pits.10131

Ballard, W. S., Tighe, P. L., & Dalton, E. F. (1991). *Examiner's manual IDEA Proficiency Test*. Brea, CA: Ballard & Tighe Publishers.

Bayley, R., & Bonnici, L. M. (2009). Recent research on Latinos in the USA and Canada, part 1: Language maintenance and shift and English Varieties. *Language and Linguistics Compass, 3/5,* 1300-1313. doi: 10.1111/j.1749-818x.2009.00159.x

Belar, C. D. (2009). Advancing the culture of competence. *Training and Education in Professional Psychology, 3,* S63-S65. doi: 10.1037/a0017541

Bialystok, E. (2011a). Coordination of executive functions in monolingual and bilingual children. *Journal of Experimental Child Psychology, 110,* 461-468. doi: 10.1016/j.jecp.2011.05.005

Bialystok, E. (2011b). Reshaping the mind: The benefits of bilingualism. *Canadian Journal of Experimental Psychology, 65,* 229-235. doi: 10.1037/a0025406

Bialystok, E., & Craik, F. I. (2010). Cognitive and linguistic processing in the bilingual mind. *Current Directions in Psychological Science, 19,* 19-23. doi: 10.1177/0963721 409358571

Bialystok, E., Craik, F. I., & Luk, G. (2012). Bilingualism: Consequences for mind and brain. *Trends in Cognitive Sciences, 16,* 240-250. doi: 10.1016/j.tics.2012.03.001

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

Brown, J. E. (2008). *The use and interpretation of the Batería III with U.S. bilinguals.* (Unpublished doctoral dissertation). Portland State University, Portland, OR.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since *Bias in mental testing*. *School Psychology Quarterly, 14,* 208-238. doi: 10.1037/h0089007

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Cabassa, L. J. (2003). Measuring acculturation: Where we are and where we need to go. *Hispanic Journal of Behavioral Sciences, 25,* 127-146. doi: 10.1177/0739986303025002001

Caltabiano, L. F. (February, 2002). *Content validation of cognitive and achievement tests through expert consensus.* Poster presented at the annual convention of the National Association of School Psychologists, Chicago, IL.

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

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

Collier, C. (1988). *Acculturation quick screen*. Ferndale, WA: Crosscultural Development Education Services.

Collier, C. (2001). *Acculturation quick screen*. Ferndale, WA: Crosscultural Development Education Services.

Cormier, D. C. (2012). *The influences of linguistic demand and cultural loading on cognitive test scores.* (Unpublished doctoral dissertation). The University of Minnestota, Minneapolis, Minnesota.

Coutinho, M. J., & Oswald, D. P. (2006). *Disproportionate representation of culturally and linguistically diverse students in special education: Measuring the problem*. Retrieved from National Center for Culturally Responsive Educational Systems website: http://www.niusileadscape.org/docs/FINAL_PRODUCTS/LearningCarousel/disproportionate_ representation.pdf

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

DeAvila, E., & Duncan, S. (2005). *Language assessment scales*. New York: McGraw-Hill.

Dhaniram-Beharry, E. (2008). *Cultural and linguistic influences on test performance: Evaluation of alternate variables.* (Unpublished doctoral dissertation). Saint John's University, Queens, New York.

Diana v. State Board of Education, C. A. 70 RFT (N. D. Cal. Feb. 3, 1970).

Edwards, O. W., & Oakland, T. D. (2006). Factorial invariance of Woodcock-Johnson III scores for African Americans and Caucasian Americans. *Journal of Psychoeducational Assessment, 24,* 358-366. doi: 10.1177/0734282906289595

Education for All Handicapped Children Act of 1975, Public Law No. 94-142, (Supp. 1984).

Educational Testing Service. (2005, July). *Comprehensive English language learning assessment: Technical summary report*. Princeton, NJ: Author.

Elliott, C. (2007). *Differential Ability Scales* (2nd ed.). San Antonio, TX: Harcourt Assessment.

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

Flanagan, D. P. (2000). Wechsler-based CHC cross-battery assessment and reading achievement: Strengthening the validity of interpretations drawn from Wechsler test scores. *School Psychology Quarterly, 15,* 295-329. doi: 10.1037/h0088789

Flanagan, D. P., & Ortiz, S. (2001). *Essentials of cross-battery assessment*. New York: Wiley.

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.

Flanagan, D. P., Ortiz, S. O., Alfonso, V. C., & Dynda, A. M. (2008). Best practices in cognitive assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 633-659). Bethesda, MD: National Association of School Psychologists.

Flanagan, D. P., Ortiz, S. O., Alfonso, V. C., & Mascolo, J. T. (2002). *The achievement test desk reference (ATDR): Comprehensive assessment and learning disabilities.* Boston: Allyn & Bacon.

Flanagan, D. P., Ortiz, S. O., Alfonso, V. C., & Mascolo, J. T. (2006). *Achievement test desk reference: A guide to learning disability identification* (2nd ed.). New York: Wiley.

Floyd, R. G., Bergeron, R., Hamilton, G., & Parra, G. R. (2010). How do executive functions fit with the Cattell-Horn-Carroll model? Some evidence from a joint factor analysis of the Delis-Kaplan Executive Function System and the Woodcock-Johnson III Tests of Cognitive Abilities. *Psychology in the Schools, 47,* 721-738. Retrieved from http://ezaccess.libraries.psu.edu/login?url=http://search.proquest.com.ezaccess.libraries.psu.edu/docview/755203993?accountid=13158

Floyd, R. G., Bergeron, R., McCormack, A. C., Anderson, J. L., & Hargrove-Owens, G. L. (2005). Are Cattell-Horn-Carroll broad ability composite scores exchangeable across batteries? *School Psychology Review, 34*,329-357. Retrieved from http://search.proquest.com.ezaccess.libraries.psu.edu/docview/219646456/fulltextPDF/1355B25CB916883E71B/1?accountid=13158

Floyd, R. G., McGrew, K. S., Barry, A., Rafael, F., & Rogers, J. (2009). General and specific effects on Cattell-Horn-Carroll broad ability composites: Analysis of the Woodcock-Johnson III normative update Cattell-Horn-Carroll factor clusters across development. *School Psychology Review, 38,* 249-265. Retrieved from http://ezaccess.libraries.psu.edu/login?url=http://search.proquest.com.ezaccess.libraries.psu.edu/docview/622034598?accountid=13158

Floyd, R. G., McGrew, K. S., & Evans, J. J. (2008). The relative contributions of the Cattell-Horn-Carroll cognitive abilities in explaining writing achievement during childhood and adolescence. *Psychology in the Schools, 45,* 132-144. doi: 10.1002/pits.20284

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist 54*, 5-20.

Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2003). Multifactored and cross-battery ability assessments: Are they worth the effort? In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence, aptitude, and achievement* (2nd ed., pp. 343-374). New York: Guilford Press.

Goddard, H. H. (1917). Mental tests and the immigrant. *The Journal of Delinquency, 2*, 243-277.

Gravois, T. A., & Rosenfield, S. A. (2006). Impact of instructional consultation teams on the disproportionate referral and placement of minority students in special education. *Remedial and Special Education, 27*, 42-52. doi: 10.1177/07419325060270010501

Green, T. D., McIntosh, A. S., Cook-Morales, V. J., & Robinson-Zañartu, C. (2005). From old schools to tomorrow's schools: Psychoeducational assessment of African American students. *Remedial and Special Education, 26*, 82-92. doi: 10.1177/07419325050260020301

Henderson, A. R. (1993). Assessing test accuracy and its clinical consequences: A primer for receiver operating characteristic curve analysis. *Annals of Clinical Biochemistry, 30,* 521-539.

Horn, J. (1999). *RE: WJR Cognitive as a measure of g?* Comments quoted from posting made to the CHC listserve, August 2, 1999.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*, 253-270. doi: 10.1037/h0023816.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi: 10.1080/10705519909540118

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

Jensen, A. R. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs, 90,* 185- 244.

Jensen, A. R. (1976). Test bias and construct validity. *Phi Delta Kappan, 58,* 340-346.

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,* 12-22. doi: 10.1037/a0030880

Kaufman, A. S. (2000). Foreword. In D. P. Flanagan, K. S. McGrew, & S. O. Ortiz, *The Wechsler Intelligence Scales and Gf-Gc Theory: A contemporary approach to interpretation.* (pp. xiv-xv). Needham Heights, MA: Allyn & Bacon.

Kaufman, A. S., Johnson, C. K., & Liu, X. (2008). A CHC theory-based analysis of age differences on cognitive abilities and academic skills at ages 22 to 90 years. *Journal of Psychoeducational Assessment, 26,* 350-381. doi: 10.1177/0734282908314108

Kaufman, A. S., & Kaufman, N. L. (2004). *Manual for the Kaufman Assessment Battery for Children - Second Edition.* Circle Pines, MN: American Guidance Service.

Keith, T. Z. (1999). Effects of general and specific abilities on student achievement: Similarities and differences across ethnic groups. *School Psychology Quarterly, 14,* 239-262. doi: 10.1037/h0089008

Keith, T. Z., Kranzler, J. H., & Flanagan, D. P. (2001). What does the Cognitive Assessment System (CAS) measure? Joint confirmatory factor analysis of the CAS and the Woodcock-Johnson Tests of Cognitive Ability (3rd edition). *School Psychology Review, 30,* 89-118. Retrieved from htt;://ezaccess.libraries.psu.edu/login?url=http://search.proquest.com.ezaccess.libraries.psu.edu/docview/619676015?accountid=13158

Keith, T. Z., & Reynolds, M. R. (2010). Cattell-Horn-Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools, 47,*635-650. doi: 10.1002/pits.20496

Kenyon, D. M. (2006). *Development and field test of ACCESS for ELLs English language proficiency test* (Technical Report #1). Retrieved from World-Class Instructional Design and Assessment (WIDA) Consortium website: http://www.wida.us/assessment/ACCESS /Tech Reports/Technical%20Report%201.pdf

Kranzler, J. H., Flores, C. G., & Coady, M. (2010). Examination of the cross-battery approach for the cognitive assessment of children and youth from diverse linguistic and cultural backgrounds. *School Psychology Review, 39,* 431-446.

Lakin, J. M. (2012). Assessing the cognitive abilities of culturally and linguistically diverse students: Predictive validity of verbal, quantitative, and nonverbal tests. *Psychology in the Schools, 49,* 756-768. doi: 10.1002/pits

Lau v. Nichols, **414 U.S. 563, 94 (1974).**

Lella Souravlis, S. A. (2010). *Evaluating speech-language and cognitive impairment patterns via the culture-language interpretive matrix.* (Unpublished doctoral dissertation). Saint John's University, Queens, New York.

Lopez, E. C., & Bursztyn, A. M. (2013). Future challenges and opportunities: Toward culturally responsive training in school psychology. *Psychology in the Schools, 50,* 212-228. doi: 10.1002/pits.21674

Lopez, E. C., Lamar, D., & Scully-Demartini, D. (1997). The cognitive assessment of limited-English-proficient children: Current problems and practical recommendations. *Cultural Diversity and Mental Health, 3,* 117-130. doi: 10.1037/1099-9809.3.2.117

Maculaitis, J. D. (2003). *Maculaitis II Test of English Language Proficiency handbook with norms tables*. (Revised ed.). Brewster, NY: Touchstone Applied Science Associates.

Marín, G., & Gamba, R. J. (1996). A new measurement of acculturation for Hispanics: The Bidimensional Acculturation Scale for Hispanics (BAS), *Hispanic Journal of Behavioral Sciences, 18*, 297-316. doi: 10.1177/07399863960183002

Matthey, S., & Petrovski, P. (2002). The Children's Depression Inventory: Error in cutoff scores for screening purposes. *Psychological Assessment, 14*, 146-149. doi: 10.1037/1040-3590.14.2.146

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50*, 215-241. doi: 10.1146/annurev.psych.50.1.215

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-GC framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151-179). New York: Guilford Press.

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

McGrew, K. S., Dailey, D. E. H., & Schrank, F. A. (2007). Woodcock-Johnson III/Woodcock-Johnson III normative update *score differences: What the user can expect and why* (Woodcock-Johnson III Assessment Service Bulletin No. 9). Rolling Meadows, IL: Riverside Publishing.

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

McGrew, K. S., & Wendling, B. J. (2010). Cattell-Horn-Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools, 47,* 651-675. Retrieved from http://ezaccess.libraries.psu.edu/login?url=http://search.proquest.com.ezaccess.libraries.psu.edu/docview/755203605?accountid=13158

Mercer, J. R. (1979). *System of multicultural pluralistic assessment: Technical manual.* New York, NY: Psychological Corporation.

Mindt, M. R., Arentoft, A., Germano, K. K., D'Aquila, E., Scheiner, D., Pizzirusso, M., Sandoval, T. C., & Gollan, T. H. (2008). Neuropsychological, cognitive, and theoretical considerations for evaluation of bilingual individuals. *Neuropsychology Review, 18*, 255-268. doi: 10.1007/s11065-008-9069-7

Muñoz-Sandoval, F. G., Woodcock, R. W., McGrew, K. S., & Mather, N. (2005). *Batería III Woodcock-Muñoz*. Itasca, IL: Riverside Publishing.

National Association of School Psychologists. (2010). *Principles for professional ethics.* Bethesda, MD: Author. Retrieved from http://www.nasponline.org/standards/2010standards/1_%20Ethical%20Principles.pdf

Newell, M. L., Nastasi, B. K., Hatzichristou, C., Jones, J. M., Schanding, G. T., & Yetter, G. (2010). Evidence on multicultural training in school psychology: Recommendations for future directions. *School Psychology Quarterly, 25,* 249-278. doi: 10.1037/a0021542

Nieves-Brull, A. I. (2006). *Evaluation of the culture-language matrix: A validation study of test performance in monolingual English speaking and bilingual English/Spanish speaking populations.* (Unpublished doctoral dissertation). Saint John's University, Queens, New York.

Ortiz, S. O. (2008). Best practices in nondiscriminatory assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 661-678). Bethesda, MD: National Association of School Psychologists.

Ortiz, S. O., & Flanagan, D. P. (2002). Some cautions concerning "Some cautions concerning cross-battery assessment." (Part I). *NASP Communiqué, 30,* 32-34. Retrieved from http://facpub.stjohns.edu/~flanagad/crossbattery/downloads/Some%20Cautions%20Concerning%20Cross-Battery%20Assessment%20Part%20I.pdf

Ortiz, S. O., & Flanagan, D. P. (2002). Some cautions concerning "Some cautions concerning cross-battery assessment." (Part II). *NASP Communiqué, 30,* 36-38. Retrieved from http://facpub.stjohns.edu/~flanagad/crossbattery/downloads/Some%20Cautions%20Concerning%20Cross-Battery%20Assessment%20Part%20II.pdf

Overton, T., Fielding, C., & Simonson, M. (2004). Decision making in determining eligibility of culturally and linguistically diverse learners: Reasons given by assessment personnel. *Journal of Learning Disabilities, 37,* 319-330. doi: 10.1177/00222194040370040401

Phelps, L., McGrew, K. S., Knopik, S. N., & Ford, L. (2005). The general (*g*), broad, and narrow CHC stratum characteristics of the WJ III and WISC-III tests: A confirmatory cross-battery investigation. *School Psychology Quarterly, 20,* 66-88. doi: 10.1521/scpq.20.1.66.64191

Piaget, J. (1962). The stages of the intellectual development of the child. *Bulletin of the Menninger Clinic, 26,* 120-128.

Plake, B. S., Impara, J. C., & Spies, R. A. (2003). Review of *language proficiency test series*. (Series Eds.), *The fifteenth mental measurement yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Prifitera, A., Weiss, L. G., & Saklofske, D. H. (1998). The WISC-III in context. In A. Prifitera & D. Saklofske (Eds.), *WISC-III clinical use and interpretation* (pp. 1-39). San Diego, CA: Academic Press.

Redfield, R., Lenton, R., & Herskovits, M. J. (1936). Memorandum for the study of acculturation. *American, Anthropologist, 38,* 149-152.

Reynolds, C. R. (2000). Why is psychometric research on bias in mental testing so often ignored? *Psychology, Public Policy, and Law, 6,* 144-150. doi: 10.1037/1076-8971.6.1.144

Rhodes, R.L., Ochoa, S.H., & Ortiz, S.O. (2005). *Assessing culturally and linguistically diverse students: A practical guide.* New York: Guilford.

Rogers, M. R., Ingraham, C. L., Bursztync, A., Cajigas-Segredo, N., Esquivel, G., Hess, R., Nahari, S. G., & Lopez, E. C. (1999). Providing psychological services to racially, ethnically, culturally, and linguistically diverse individuals in the schools: Recommendations for practice. *School Psychology International, 20,* 243-264. doi: 10.1177/0143034399203001

Rogers-Adkinson, D. L., Ochoa, T. A., & Delgado, B. (2003). Developing cross-cultural competence: Serving families of children with significant developmental needs. *Focus on Autism and Other Developmental Disablities, 18,* 4-8.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition (SB-V).* Itasca, IL: Riverside Publishing.

Rueda, R., & Windmueller, M. P. (2006). English language learners, LD, and overrepresentation. *Journal of Learning Disabilities, 39,* 99-107. doi: 10.1177/00222194060390020801

Salvia, J., & Ysseldyke, J. (1991). *Assessment in special and remedial education* (5th Ed.) Boston, MA: Houghton-Mifflin.

Sánchez, G. I. (1934). Bilingualism and mental measures: A word of caution. *Journal of Applied Psychology, 18,* 765-772. doi: 10.1037/h0072798

Sanchez, S. V., Rodriguez, B. J., Soto-Huerta, M. E., Villereal, F. C., & Guerra, N. S. (2013). A case for multidimensional bilingual assessment. *Language Assessment Quarterly, 10,* 160-177.

Sanders, S., McIntosh, D. E., Dunham, M., Rothlisberg, B. A., & Finch, H. (2007). Joint confirmatory factor analysis of the Differential Ability Scales and the Woodcock-Johnson Tests of Cognitive Abilities-Third Edition. *Psychology in the Schools, 44,* 119-138. doi: 10.1002/pits.20211

Sandoval, J., Frisby, C. L., Geisinger, K. F., Scheuneman, J. D., & Grenier, J. R. (Eds.). (1998). *Test interpretation and diversity: Achieving equity in assessment.* Washington, DC: American Psychological Association.

Sattler, J. M. *Assessment of children: Cognitive applications, Fourth edition.* La Mesa, CA: Jerome M. Sattler, Publisher, Inc.

Schrank, F. A., McGrew, K. S., & Woodcock, R. W. (2001). *Technical abstract* (Woodcock-Johnson III Assessment Service Bulletin No. 2). Itasca, IL: Riverside Publishing.

Skutnabb-Kangas, T. (2007). Linguistic human rights in education? In O. García & C. Baker (Eds.), *Bilingual education: An introductory reader* (pp. 137-144). Retrieved from http://books.google.com/books?hl=en&lr=&id=PQoVx1dlXwC&oi=fnd&pg=PA137&dq=define+bilingual&ots=5FhXks9z6t&sig=r9JdZLsCJHs1CYaK1C9H5oXVFEk#v=onepage&q=define%20bilingual&f=false

Sotelo-Dynega, M. (2008). *Cognitive performance and the development of English language proficiency.* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3282715)

Styck, K. (2012). *Diagnostic utility of the Culture-Language Interpretive Matrix for the WISC-IV among referred students.* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3517798)

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293. doi: 10.1126/science.3287615.

Taub, G. E., & McGrew, K. S. (2004). A confirmatory factor analysis of Cattell-Horn-Carroll theory and cross-age invariance of the Woodcock-Johnson Tests of Cognitive Abilities III. *School Psychology Quarterly, 19,* 72-87. doi: 10.1521/scpq.19.1.72.29409

Templeton, M. M. (2012). *An examination of the effects of culture and language on the executive functioning of Spanish-speaking English learners according to the Delis-Kaplan Executive Function System.* (Unpublished doctoral dissertation). Alliant International University, San Diego, California.

Tychanska, J. (2009). *Evaluation of speech and language impairment using the culture-language test classifications and interpretive matrix.* (Unpublished doctoral dissertation). Saint John's University, Queens, New York.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias.* Norwood, NJ: Ablex.

Verderosa, F. A. (2007). *Examining the effects of language and culture on the Differential Ability Scales with bilingual preschoolers.* (Unpublished doctoral dissertation). Saint John's University, Queens, New York.

Wallace, B. J. (2000). A call for change in multicultural training at graduate schools of education: Educating to end oppression for social justice. *Teachers College Record, 102,* 1086-1111.

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997). Discriminant and predictive validity of the WISC-III ACID profile among children with learning disabilities. *Psychology in the Schools, 34,* 309-319.

Watkins, M. W., Youngstrom, E. A., & Glutting, J, J. (2002). Some cautions concerning cross-battery assessment. *NASP Communiqué, 30,* 16-19.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children, Third Edition (WISC-III).* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale, Third Edition (WAIS-III).* San Antonio, TX: Harcourt Assessment.

Wechsler, D. (2002). *Wechsler Primary and Preschool Scale of Intelligence, Third edition (WPPS-III).* San Antonio, TX: Harcourt Assessment.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV).* San Antonio, TX: Harcourt Assessment.

Woodcock, R. W., McGrew, K. S., Schrank, F. A., & Mather, N. (2007). *Woodcock-Johnson III normative update.* Rolling Meadows, IL: Riverside.

APPENDIX

Table 42

Means and Standard Deviations of WJ-III-NU Subtests for ELLs Not Identified with SLD from the Unmatched Sample

Degree of Cultural Loading	Degree of Linguistic Demand: Low		Moderate		High	
	Test Name	***M (SD)***	**Test Name**	***M (SD)***	**Test Name**	***M (SD)***
Low	Spatial Relations	99.4 (9.2)	Numbers Reversed	89.2 (14.2)	Analysis-Synthesis	100.2 (13.3)
			Visual Matching	86.9 (15.7)	Concept Formation	93.9 (13.0)
	Cell Average	**99.4 (9.2)**	**Cell Average**	**88.1 (15.0)**	**Cell Average**	**97.1 (13.2)**
Moderate	Picture Recognition	104.1 (11.0)	Retrieval Fluency	90.1 (15.2)	Auditory Attention	97.9 (11.5)
			Visual Auditory Learning	91.1 (10.9)	Decision Speed	100.6 (15.4)
					Memory for Words	88.7 (14.3)
					Sound Blending	97.4 (12.0)
	Cell Average	**104.1 (11.0)**	**Cell Average**	**90.6 (13.1)**	**Cell Average**	**96.2 (13.3)**
High					General Information	92.5 (15.6)
					Verbal Comprehension	90.6 (11.3)
	Cell Average		**Cell Average**		**Cell Average**	**91.6 (13.5)**

Note. n = 40. WJ-III-NU = Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update, ELL =English language learner, SLD = Specific Learning Disability.

Table 43

Means and Standard Deviations of WJ-III-NU Subtests for ELLs Identified with SLD from the Unmatched Sample

Degree of Cultural Loading	Degree of Linguistic Demand: Low		Moderate		High	
	Test Name	***M (SD)***	**Test Name**	***M (SD)***	**Test Name**	***M (SD)***
Low	Spatial Relations	98.0 (7.8)	Numbers Reversed	88.3 (15.3)	Analysis-Synthesis	95.6 (12.5)
			Visual Matching	82.3 (14.9)	Concept Formation	89.4 (11.8)
	Cell Average	**98.0 (7.8)**	**Cell Average**	**85.3 (15.1)**	**Cell Average**	**92.5 (12.2)**
Moderate	Picture Recognition	102.6 (12.8)	Retrieval Fluency	85.9 (16.9)	Auditory Attention	97.5 (17.7)
			Visual Auditory Learning	87.1 (11.1)	Decision Speed	94.4 (16.3)
					Memory for Words	83.3 (12.4)
					Sound Blending	94.8 (12.4)
	Cell Average	**102.6 (12.8)**	**Cell Average**	**86.5 (14.0)**	**Cell Average**	**92.6 (14.7)**
High					General Information	82.4 (13.1)
					Verbal Comprehension	85.1 (9.6)
	Cell Average		**Cell Average**		**Cell Average**	**83.8 (11.4)**

Note. n = 78. WJ-III-NU = Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update; ELL =English language learner, SLD = Specific Learning Disability.

Table 44

Results of Independent Samples t tests on WJ-III-NU Scores for ELLs based on SLD Status

WJ-III-NU Test	Non-SLD (n = 40)	SLD (n = 41)	p value	d
Verbal Comprehension	90.60	85.10	.020	0.53
General Information	92.53	82.41	.002	0.71
Concept Formation	93.90	89.41	.108	0.37
Analysis-Synthesis	100.15	95.59	.116	0.36
Visual Auditory Learning	91.13	87.12	.106	0.37
Retrieval Fluency	90.10	85.85	.239	0.27
Spatial Relations	99.35	97.98	.470	0.16
Picture Recognition	104.10	102.56	.564	0.13
Sound Blending	97.43	94.78	.333	0.22
Auditory Attention*	97.93	97.46	.890	0.04
Visual Matching	86.85	82.27	.181	0.30
Decision Speed	100.55	94.41	.086	0.39
Numbers Reversed	89.18	88.32	.795	0.06
Memory for Words	88.65	83.29	.076	0.41

Note. N = 81. ELL = English language learner, SLD = specific learning disability. WJ-III-NU = Woodcock-Johnson Tests of Cognitive Ability – Third Edition, Normative Update.

* Levene's Test for Equality of Variances was significant (p = .006) so a t-test with equal variances not assumed was interpreted.

Table 45

Template of C-LIM Diagnostic Decisions for ELL Students Based on SLD Status

		C-LIM Decisions	
		Declining Pattern	No Declining Pattern
SLD Status	Non-SLD	TP	FN
	SLD	FP	TN

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, SLD = Specific Learning Disability, TP = true positive, FP = false positive, FN = false negative, TN = true negative.

Table 46

Frequency Count of C-LIM Decision for the Unmatched ELL Sample Based on SLD Status for Influence of Cultural Loading Using the Most Stringent Interpretation

		Most Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
SLD Status	Non-SLD (*n*)	0	40	40
	SLD (*n*)	0	41	41
	Total	0	81	81

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, SLD = specific learning disability.

Table 47

Frequency Count of C-LIM Decision for the Unmatched ELL Sample Based on SLD Status for Influence of Linguistic Demand Using the Most Stringent Interpretation

		Most Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
SLD Status	Non-SLD (*n*)	0	40	40
	SLD (*n*)	0	41	41
	Total	0	81	81

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, SLD = specific learning disability.

Table 48

Frequency Count of C-LIM Decision for the Unmatched ELL Sample Based on SLD Status for Influence of Cultural Loading and Linguistic Demand Using the Most Stringent Interpretation

		Most Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
SLD Status	Non-SLD (*n*)	0	40	40
	SLD (*n*)	0	41	41
	Total	0	81	81

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, SLD = specific learning disability.

Table 49

Frequency Count of C-LIM Decision for the Unmatched ELL Sample Based on SLD Status for Influence of Cultural Loading Using the Moderately Stringent Interpretation

		Moderately Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
SLD Status	Non-SLD (*n*)	6	34	40
	SLD (*n*)	8	33	41
	Total	14	67	81

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, SLD = specific learning disability.

Table 50

Frequency Count of C-LIM Decision for the Unmatched ELL Sample Based on SLD Status for Influence of Linguistic Demand Using the Moderately Stringent Interpretation

		Moderately Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
SLD Status	Non-SLD (*n*)	2	38	40
	SLD (*n*)	5	36	41
	Total	7	74	81

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, SLD = specific learning disability.

Table 51

Frequency Count of C-LIM Decision for the Unmatched ELL Sample Based on SLD Status for Influence of Cultural Loading and Linguistic Demand Using the Moderately Stringent Interpretation

		Moderately Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
SLD Status	Non-SLD (*n*)	1	39	40
	SLD (*n*)	3	38	41
	Total	4	77	81

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, SLD = specific learning disability.

Table 52

Frequency Count of C-LIM Decision for the Unmatched ELL Sample Based on SLD Status for Influence of Cultural Loading Using the Least Stringent Interpretation

		Least Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
SLD Status	Non-SLD (*n*)	12	28	40
	SLD (*n*)	9	32	41
	Total	21	60	81

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, SLD = specific learning disability.

Table 53

Frequency Count of C-LIM Decision for the Unmatched ELL Sample Based on SLD Status for Influence of Linguistic Demand Using the Least Stringent Interpretation

		Least Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
SLD Status	Non-SLD (*n*)	3	37	40
	SLD (*n*)	7	34	41
	Total	10	71	81

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, SLD = specific learning disability.

Table 54

Frequency Count of C-LIM Decision for the Unmatched ELL Sample Based on SLD Status for Influence of Cultural Loading and Linguistic Demand Using the Least Stringent Interpretation

		Least Stringent C-LIM Decisions		
		Declining Pattern (*n*)	No Declining Pattern (*n*)	Total
SLD Status	Non-SLD (*n*)	10	30	40
	SLD (*n*)	18	23	41
	Total	28	53	81

Note. C-LIM = Culture-Language Interpretive Matrix, ELL = English language learner, SLD = specific learning disability.

Table 55

Summary of Frequencies of ELLs with and without SLD Patterns of Decline Based on the Three C-LIM interpretations and Age of Entry into the U.S.

Pattern of Decline		Born in U.S.				Age 0.1 to 3				Age 3.1 to 7				Age 7.1 to 12			
		Non-SLD		SLD		Non-SLD		SLD		Non-SLD		SLD		Non-SLD		SLD	
		Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)	Yes (*n*)	No (*n*)
Most Stringent	Cultural	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Linguistic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Combined	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Moderately Stringent	Cultural	5	27	6	31	0	2	1	1	0	1	0	1	1	1	0	0
	Linguistic	1	31	5	32	0	2	0	2	1	0	0	1	0	2	0	0
	Combined	1	31	2	35	0	2	1	1	0	1	0	1	0	2	0	0
Least Stringent	Cultural	12	20	8	29	0	2	0	2	0	1	0	1	0	2	0	0
	Linguistic	3	29	7	30	0	2	0	2	0	1	0	1	0	2	0	0
	Combined	9	23	17	20	0	2	1	1	0	1	0	1	0	2	0	0

Note. $N = 77$. ELL = English language learner, SLD = specific learning disability, C-LIM = Culture-Language Interpretive Matrix.

Table 56

Summary of Frequencies of ELLs' Language Proficiency Level Based on Patterns of Decline and the Three Levels of C-LIM Interpretation

WIDA Score		Level 1				Level 2				Level 3				Level 4				Level 5				Level 6			
		Non-SLD		SLD		Non-SLD		SLD		Non-SLD		SLD		Non-SLD		SLD		Non-SLD		SLD		Non-SLD		SLD	
Pattern of Decline		Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Most Stringent	Cultural	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Linguistic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Combined	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Moderately Stringent	Cultural	0	2	0	0	0	1	2	2	3	10	4	14	1	5	0	7	0	4	0	1	0	0	0	0
	Linguistic	0	2	0	0	0	1	0	4	1	12	3	15	0	6	2	5	0	4	0	1	0	0	0	0
	Combined	0	2	0	0	0	1	1	3	0	13	0	18	0	6	0	7	0	4	0	1	0	0	0	0
Least Stringent	Cultural	1	1	0	0	1	0	2	2	3	10	4	14	2	4	2	5	1	3	0	1	0	0	0	0
	Linguistic	0	2	0	0	0	1	0	4	0	13	4	14	2	4	2	5	0	4	0	1	0	0	0	0
	Combined	0	2	0	0	0	1	2	2	6	7	10	8	0	6	3	4	0	4	0	1	0	0	0	0

Note. N = 56. WIDA = World-Class Instructional Design and Assessment, ELL = English language learner, C-LIM = Culture-Language Interpretive Matrix.

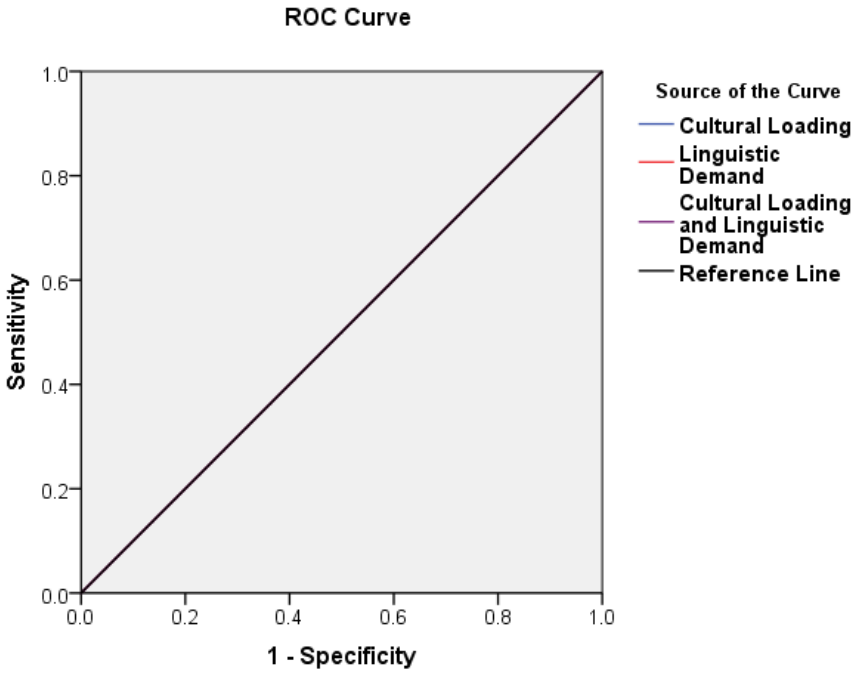

Figure 12. ROC curves for ELLs with and without SLD based on all three types of decline in WJ-III-NU scores (cultural only, linguistic only, and combined) using the most stringent interpretation. ELLs with SLD = 41; ELLs without SLD = 40.

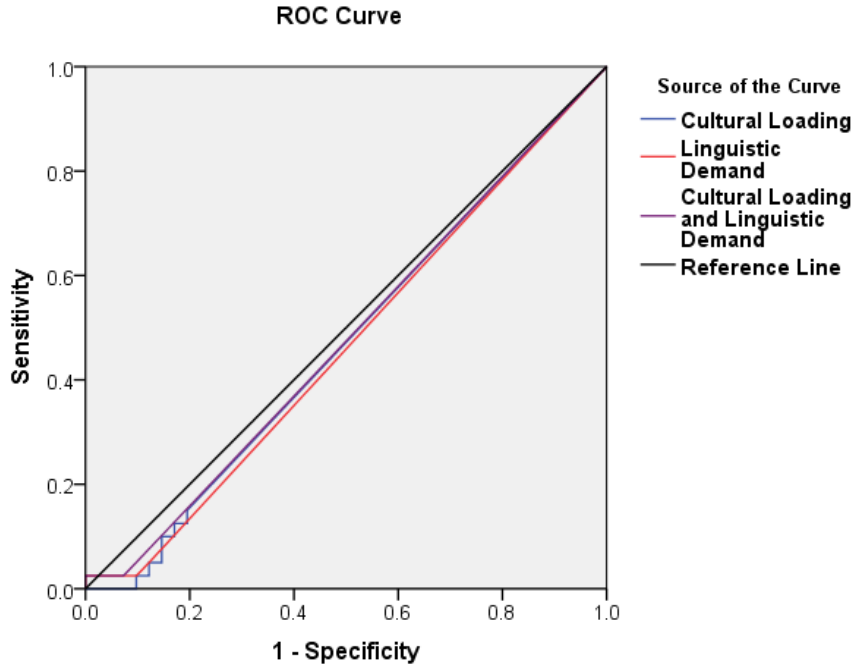

Figure 13. ROC curves for ELLs with and without SLD based on all three types of decline in WJ-III-NU scores (cultural only, linguistic only, and combined) using the moderately stringent interpretation. ELLs with SLD = 41; ELLs without SLD = 40.

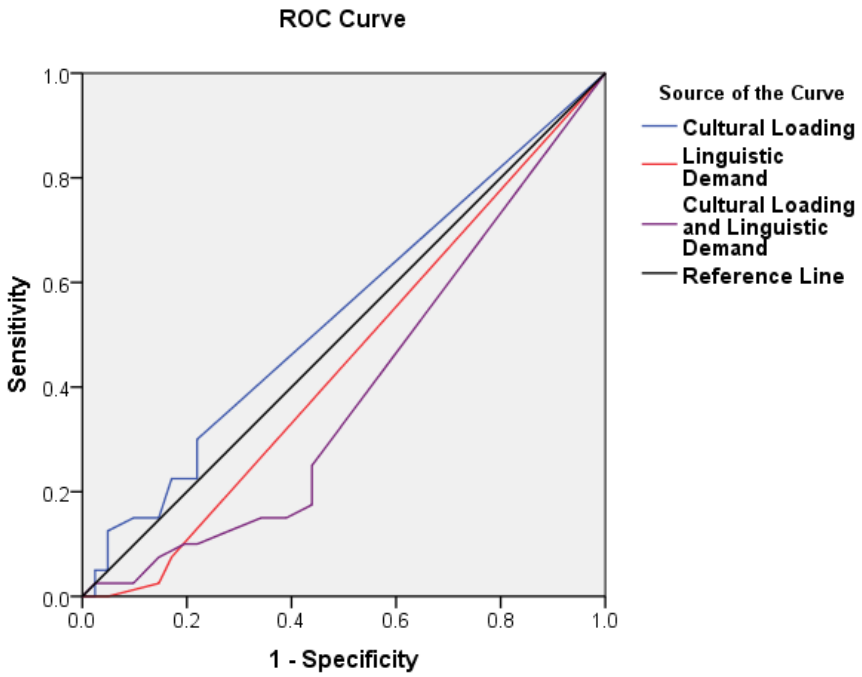

Figure 14. ROC curves for ELLs with and without SLD based on all three types of decline in WJ-III-NU scores (cultural only, linguistic only, and combined) using the least stringent interpretation. ELLs with SLD = 41; ELLs without SLD = 40.

Table 57

Summary of Diagnostic Utility Statistics for ELLs with or without SLD in the Unmatched Sample Based on All levels of C-LIM Interpretation Criteria

	Pattern of Decline	Sensitivity Value	Specificity Value	PPV Value	NPV Value	AUC
Most Stringent	Cultural Loading	0.00	1.00	undefined	0.51	.50
	Linguistic Demand	0.00	1.00	undefined	0.51	.50
	Combined	0.00	1.00	undefined	0.51	.50
Moderately Stringent	Cultural Loading	0.15	0.80	0.43	0.49	.47
	Linguistic Demand	0.05	0.88	0.29	0.49	.46
	Combined	0.03	0.93	0.25	0.49	.48
Least Stringent	Cultural Loading	0.30	0.78	0.57	0.53	.54
	Linguistic Demand	0.08	0.83	0.30	0.48	.45
	Combined	0.25	0.56	0.36	0.43	.39

Note. *N* = 81; PPV = positive predictive value, NPV = negative predictive value, AUC = area under curve.

Table 58

Binary AUC Values and Percent of ELLs with a Declining Pattern of Scores based on Moderately Stringent Criteria

	Cultural Loading Only			Linguistic Demand Only			Cultural Loading & Linguistic Demand		
Cut Score	Non-SLD %	SLD %	AUC	Non-SLD %	SLD %	AUC	Non-SLD %	SLD %	AUC
1	15.0	19.5	.48	5.0	12.2	.46	2.5	7.3	.48
2	10.0	14.6	.48	2.5	7.3	.48	2.5	0.0	.51
3	7.5	14.6	.46	2.5	4.9	.50	0.0	0.0	
4	5.0	12.1	.46	0.0	0.0		0.0	0.0	
5	2.5	9.8	.46	2.5	2.4	.50	0.0	0.0	
6	0.0	9.8	.45	0.0	0.0		0.0	0.0	
7	0.0	4.9	.48	0.0	0.0		0.0	0.0	
8	0.0	0.0		0.0	0.0		0.0	0.0	
9	0.0	2.4	.49	0.0	0.0		0.0	0.0	
10	0.0	0.0		0.0	0.0		0.0	0.0	
11	0.0	0.0		0.0	0.0		0.0	0.0	
12	0.0	0.0		0.0	0.0		0.0	0.0	
13	0.0	0.0		0.0	0.0		0.0	0.0	

Note. N = 81, ELLs with SLD = 41, ELLs without SLD = 40. ELL = English language learner, SLD = specific learning disability, AUC = area under curve.

Table 59

Binary AUC Values and Percent of ELLs with a Declining Pattern of Scores based on Least Stringent Criteria

	Cultural Loading Only			Linguistic Demand Only			Cultural Loading & Linguistic Demand		
Cut Score	Non-SLD %	SLD %	AUC	Non-SLD %	SLD %	AUC	Non-SLD %	SLD %	AUC
1	30.0	22.0	.54	7.5	17.1	.45	25.0	43.9	.41
2	22.5	22.0	.50	2.5	14.6	.44	0.0	0.0	
3	22.5	17.1	.53	0.0	4.9	.48	17.5	43.9	.37
4	15.0	14.6	.50	0.0	0.0		15.0	39.0	.38
5	0.0	0.0		0.0	0.0		15.0	34.1	.40
6	15.0	9.8	.53	0.0	2.4	.49	10.0	22.0	.44
7	12.5	4.9	.54	0.0	0.0		10.0	19.5	.45
8	0.0	0.0		0.0	0.0		7.5	14.6	.46
9	0.0	0.0		0.0	0.0		5.0	12.2	.46
10	0.0	0.0		0.0	0.0		0.0	0.0	
11	10.0	4.9	.53	0.0	0.0		2.5	9.8	.46
12	0.0	0.0		0.0	0.0		2.5	7.3	.48
13	7.5	4.9	.51	0.0	0.0		2.5	2.4	.50
14	0.0	0.0		0.0	0.0		0.0	0.0	
15	0.0	0.0		0.0	0.0		0.0	0.0	
16	0.0	0.0		0.0	0.0		0.0	0.0	
17	5.0	4.9	.50	0.0	0.0		0.0	0.0	
18	0.0	0.0		0.0	0.0		0.0	0.0	
19	0.0	0.0		0.0	0.0		0.0	0.0	
20	0.0	0.0		0.0	0.0		0.0	0.0	
21	0.0	0.0		0.0	0.0		0.0	0.0	
22	0.0	0.0		0.0	0.0		0.0	0.0	

Note. N = 81, ELLs with SLD = 41, ELLs without SLD = 40. ELL = English language learner, SLD = specific learning disability, AUC = area under curve.

Table 59 (continued)

	Cultural Loading Only			Linguistic Demand Only			Cultural Loading & Linguistic Demand		
Cut Score	Non-SLD %	SLD %	AUC	Non-SLD %	SLD %	AUC	Non-SLD %	SLD %	AUC
23	0.0	0.0		0.0	0.0		0.0	0.0	
24	0.0	0.0		0.0	0.0		0.0	0.0	
25	5.0	2.4	.51	0.0	0.0		0.0	0.0	
26	0.0	0.0		0.0	0.0		0.0	0.0	
27	0.0	0.0		0.0	0.0		0.0	0.0	
28	0.0	0.0		0.0	0.0		0.0	0.0	
29	0.0	0.0		0.0	0.0		0.0	0.0	
30	0.0	0.0		0.0	0.0		0.0	0.0	
31	0.0	0.0		0.0	0.0		0.0	0.0	
32	0.0	0.0		0.0	0.0		0.0	0.0	
33	0.0	0.0		0.0	0.0		0.0	0.0	
34	0.0	0.0		0.0	0.0		0.0	0.0	
35	0.0	0.0		0.0	0.0		0.0	0.0	
36	2.5	2.4	.50	0.0	0.0		0.0	0.0	
…	0.0	0.0		0.0	0.0		0.0	0.0	
67	0.0	2.4	.49	0.0	0.0		0.0	0.0	

Note. N = 81, ELLs with SLD = 41, ELLs without SLD = 40. ELL = English language learner, SLD = specific learning disability, AUC = area under curve.

Curriculum Vita

Erin L. Meyer

EDUCATION

Ph.D.	**School Psychology**, Pennsylvania State University, University Park, PA, 2013 **Specialization in Culture and Language Education**
M.Ed.	**School Psychology**, Pennsylvania State University, University Park, PA 2007
B.S.	**Psychology**, Juniata College, Huntingdon, PA, 2004

PROFESSIONAL CREDENTIALS

Certified School Psychologist – Pennsylvania Department of Education
Pupil Personnel Services License in School Psychology – Virginia Department of Education

PROFESSIONAL EXPERIENCE

School Psychologist, Prince William County Public Schools, Prince William County, VA (2013 – Present)
Contract Psychologist, Loudoun County Public Schools, Loudoun County, VA (2010 – 2013)
School Psychology Intern, Loudoun County Public Schools, Loudoun County, VA (2009 – 2010)
Milieu Specialist, The Kellar School, Fairfax, VA (2008 – 2009)
Specialization in Culture and Language Education Practicum Student, State College School District, State College, PA (2007 – 2008)
Student Supervisor, Pennsylvania State University CEDAR Clinic, University Park, PA (2007 – 2008)
School Psychology Clinician, Pennsylvania State University CEDAR Clinic, University Park, PA (2005 – 2008)
Graduate Assistant, Pennsylvania State University, University Park, PA (2004-2006)

AFFILIATIONS

American Psychological Association – Division 16 (2005 – Present)
National Association of School Psychologists (2006 – Present)
Psi Chi Honor Society (2003 – Present)
Virginia Academy of School Psychologists (2012 – Present)
Virginia Psychological Association (2012 – Present)

PUBLICATIONS & PRESENTATIONS

Meyer, E., Vandiver, B., & Applegate, H. (2013, August). Diagnostic accuracy and efficiency of the Culture-Language Interpretive Matrix with the WJ-III-NU. In B. J. Vandiver (Chair) and G. L. Canivez (Discussant), *Clinical usefulness of Cultural Linguistic Interpretive Matrix with English language learners.* Symposium conducted at the annual convention of the American Psychological Association, Honolulu, HI.

Meyer, E., Schaefer, B., Merino Soto, C., Simmons, C., Anguiano, R., Worrell, F., Brett, J., Holman, A., Martin, J., Hata, H., Roberts, K., & Mello, Z. (2011). Factor Structure of Child Behavior Scale scores in Peruvian preschoolers. *Psychology in the Schools, 48*, 931-942. doi: 10.1002/pits20596.

Reid, E. E., Goffreda, C. T., Culler, E. D., McGinnis, A. M., Miller, A. R., Reid, M. A., Freberg., M. E., Meyer, E. L., & Hahn, K. R. (2009, February). *Construct validity of the WJ-III Cognitive among adjudicated adolescents.* Poster presentation at the annual convention of the National Association of School Psychologists, Boston, MA.