The Influences of Linguistic Demand and Cultural Loading on Cognitive Test Scores

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Damien Clement Cormier

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Dr. James Ysseldyke

June, 2012

Acknowledgements

I would like to thank my adviser, Dr. James Ysseldyke, for his encouragement and support throughout my graduate career. The valuable knowledge that he has conveyed over the years has not only helped me overcome obstacles along the way, but will also serve me well over the course of my career in academia.

I would also like to thank my friends and family for their support. I would like to dedicate my work to my parents who have always encouraged me to accomplish great things. C'est grâce à vous que j'ai les moyens d'atteindre mes buts ambitieux!

Finally, I would like to thank the members of my dissertation committee, Drs. Kevin McGrew, David Johnson, and Annie Hansen, for their support of this research, as well as their valuable knowledge and guidance. I would like to give special recognition to Dr. McGrew for his outstanding mentoring, which has allowed me to further develop my abilities as a researcher.

Abstract

The increased diversity of the student population in the United States has resulted in increased concerns related to the assessment of cognitive abilities for students from culturally and linguistically diverse backgrounds. To date, little empirical exists to support recommendations in test selection and interpretation, such as those presented in the Culture-Language Interpretative Matrix (C-LIM). In this study, the researcher aimed to provide empirical evidence for the classifications from the C-LIM for the Woodcock-Johnson Test of Cognitive Abilities, Third Edition (WJ-III). The researcher used the WJ-III normative update sample to determine the extent to which the two dimensions of the C-LIM (e.g. Cultural Loading and Linguistic Demand) influence performance on the 20 tests included in the battery. A secondary set of analyses examined the individual contributions of expressive and receptive language on test performance. Results provide partial support for the dimensions of the C-LIM (e.g. Cultural Loading and Linguistic Demand). However, it appears that a re-categorization of the individual tests within the matrix is necessary. The influence of both culture and language and the practical application of the results are discussed.

Table of Contents

List of Tables

List of Figures

Chapter 1 - Introduction

The use of cognitive assessment batteries is at the heart of the debate of providing an appropriate education for students from culturally and linguistically diverse (CLD) backgrounds. The elimination of potential sources of bias during cognitive assessment is central to demonstrating that these methods are valid for high-stakes educational decisions. Assessment tools used by school psychologists often come under scrutiny when their *consequential validity* is brought into question. Consequential validity "emerged from concerns related to the societal ramifications of assessments" (Mertens, 2003, p.178). For school psychologists, social ramifications involved the the high stakes decisions (e.g., special education placement) that are made based on the results from the assessment of cognitive abilities.

Disproportionality. Disproportionality is an over- or under-representation a particular demographic group in special education programs relative to the presence of this group in the overall student population. Thus, disproportionality does not only apply to situations in which students are over-represented in special education services, but should also be considered in terms of an under-representation. For example, an under-representation of minority students in gifted and talented programs may suggest the potential for different forms of bias that are preventing students with a level of ability that is appropriate for these services from receiving them. Nonetheless, disproportionality is often confounded by states use of unreliable discrepancy models when making high stakes decisions. Despite the body of evidence that exists discouraging the use of discrepancy models in making eligibility decisions, consequential validity comes into play and measures of cognitive abilities are to blame for the disproportionate representation of minority groups receiving special education services.

Considering psychometric bias as the sole cause of disproportionality is an over-simplification. Measures of cognitive abilities are among the most theoretically rigorous and psychometrically sound measures of psychological constructs available today (McGrew, Keith, Flanagan, & Vanderwood, 1997; McGrew, 2005). Therefore, when gathering validity evidence for a particular assessment tool, test developers and researchers must take a comprehensive and multi-faceted approach to determining the extent to which a variety of biases, other than those related to the psychometric properties of the instrument, are related to the complex issue of disproportionality.

Culturally and Linguistically Diverse (CLD) Students

The term culturally and linguistically diverse has been used interchangeably with other labels such as limited English proficient, language minority, or English language learners (Garcia & Cuellar, 2006). CLD is used here due to its clear inclusion of a broader number of people, as it includes all individuals from diverse backgrounds, regardless of their English proficiency. Moreover, culture and language may provide distinctive qualities in an individual's ability to function as a part of American society.

There are close to 47 million people in the United States, over the age of 5, who speak a language other than English at home (U.S. Census, 2002). Moreover, nearly 40% of students in U.S. public schools belong to an ethnic minority group (Garcia & Cuellar, 2006). Thus, it is unsurprising that there is growing concern about how CLD individuals will shape the future of education and society in the United States. Furthermore, according to census data taken in 2000, there are across the United States close to 47 million people aged 5 years or older who speak a language other than English at home. Unfortunately, as has been the case for over 30 years, the over-representation of CLD groups in special education placements (Citation) continues to be an important issue for educators. Specifically, Klingner and colleagues (2005) suggest that measures of cognitive abilities should continue to be improved to take into account social and cultural learning or they should not be used at all in informing eligibility decisions.

Despite the increased awareness of issues related to bias in cognitive assessment of CLD students by professional associations such as those described in the *Standards for Educational and Psychological Testing* (American Educational Research Association, 2004) and the *School Psychology: A Blueprint for Training and Practice* (Ysseldyke et al., 2006), school psychologists do not feel comfortable with the interpretations that are drawn when assessing CLD students, which they attribute primarily to a lack of training in this area (Rhodes, Ochoa, & Ortiz, 2005). This finding, although alarming, suggests researchers must develop ways to reduce this effect by providing school psychologists with adequate tools to guide their interpretation of commonly used assessment procedures. Therefore, although biases cannot be totally avoided when constructing and appropriately norming a test to be used across the country, an empirically-validated indicator of the extent to which each test is affected by cultural loading and Linguistic

Demand would be extremely useful to practitioners as they attempt to make recommendations based on testing results.

The Assessment of Cognitive Abilities

Researchers have proposed a number of definitions and interpretations of intelligence over the past 100 years. These definitions range from broad statements such as: "the capacity to learn or profit by experience" to an integration of specific abilities such as "sensation, perception, association, memory, imagination, discrimination, judgment, and reasoning" (Wasserman & Tulsky, 2005, p.15). Wechsler (1939) offered perhaps the most concise, yet comprehensive definition of intelligence:

> Intelligence is the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment. It is global because it characterizes the individual's behavior as a whole; it is an aggregate because it is composed of elements or abilities which, though not entirely independent, are qualitatively differentiable. (Wechsler, 1939, p. 3)

Researchers have yet to reach a clear consensus on what constitutes intelligence, which leads to concerns about how we can measure something so broadly defined. As is often the case with this type of dilemma, theory guides our understanding of abstract constructs, with intelligence perhaps being one of the greatest examples.

Theory. The ubiquitous Cattell-Horn-Carroll (CHC) theory of cognitive abilities originated from the Cattell-Horn Gf-Gc theory and Carroll's work that used factor analysis to provide a practical application of the Gf-Gc theory (McGrew, 2005). The

CHC model is comprised of three hierarchical strata. The highest level, Stratum III, represents the composite factor of General Intelligence (*g)*. Stratum II makes the distinction between a number of broad abilities including Fluid Reasoning, Crystallized Intelligence, Short-Term Memory, Visual Processing, Auditory Processing, Long-Term Retrieval, Processing Speed, Reading and Writing Ability, Quantitative Knowledge, and Reaction Time or Decision Speed, whereas Stratum I provides a variety of narrow abilities (McGrew, 2009). As research on cognitive abilities continues to move forward, McGrew (2009) suggested that CHC theory should be used to describe the taxonomy of intelligence to facilitate understanding within the research community, as well as in practice. Although CHC theory remains central to current assessment batteries, new theories have been proposed. However, as McGrew (2009) suggested CHC theory is highly influential in the development of these new theories, such as the Verbal–Perceptual-Rotation model.

Behaviors Sampled by Intelligence Tests. The aforementioned theoretical framework is represented through a number of behaviors that are sampled during the administration of an intelligence test. Specifically, measures of intelligence sample behaviors such as discrimination, generalization, motor behavior, general knowledge, vocabulary, induction, comprehension, sequencing, detail recognition, analogical reasoning, pattern completion, abstract reasoning, and memory (Salvia, Ysseldyke, & Bolt, 2013). The ability to perform these behaviors, and the dynamics between them, changes over time and between individuals. In other words, although these are the fundamental behaviors that are sampled by intelligence tests, there is variability within the individual in the use of these behaviors when giving an item from a test battery. For example, *generalization* involves the presentation of a stimulus, which the student has to either respond to through simple or abstract matching. The content of the stimulus involved in generalization tasks may be figural, symbolic, or semantic (Salvia et al., 2013).

Application. The use of cognitive measures originated from the need to determine which individuals possessed the abilities to benefit from formal schooling (Glaser, 1981). Today, “individually administered intelligence tests are most frequently used for making exceptionality, eligibility, and educational placement decisions” (Salvia & Ysseldyke, 2004, p.204). Moreover, the access to certain educational placement still remains under the control of being determined by an assessment procedure, which may or may not include the use of a cognitive measure. One of the models still in use today is the achievement-ability discrepancy model, which contends that a student is thought to have a learning disability if he or she demonstrates a significant difference between cognitive ability (i.e. IQ) and achievement.

There are a number of problems with discrepancy models that are based on measures of IQ, such as differences in technical adequacy between tests, variable definitions of cut-off scores, and age-based criteria, which are all used in some cases to justify the label of a learning disability (Francis et al., 1991). Researchers have, however, emphasized that discrepancy scores are inappropriate for educational placement decisions (Fletcher et al., 1992; Francis et al. 1996).

Acculturation. Anthropologists have defined acculturation as a bidirectional, dynamic process that involves the interchange between two cultural groups that come into contact with each other (Sayegh & Lasry, 1993). In an educational setting, acculturation could be defined as the process of exposure and assimilation to American culture. The process of acculturation is different for each culture, although it appears to be largely multidimensional across cultures and there are a number of measures available that attempt to capture the degree of acculturation of a diverse group of individuals (Matsudaira, 2006). Acculturation is largely dependent on continuous contact occurring in order for it to progress (Padilla & Perez, 2003).

The process of acculturation influences a student's performance on measures of cognitive abilities, as these measures are inextricably linked to the language and culture in which they are developed and normed (do you need a reference here). There are three primary factors that influence English language proficiency, which could also apply to the general process of acculturation. They are the age that the student began to learn English, the context that the English language exposure occurred, and the similarity between the student's native language and English (Salvia et al., 2013). An example of how this process unfolds with language is that it generally takes 2 years to develop basic interpersonal communication skills (BICS) and at least 5 to 6 to develop their cognitive academic language proficiency (CALP; Salvia et al., 2013). However, the 3 aforementioned factors will influence the rate of progression from BICS to CALP. Thus, it becomes evident that the process of acculturation is an important consideration when testing a student from a CLD background. Practitioners must take these three factors into account, instead of considering ELL status or time in the U.S. individually, in order to make a sound judgment on the influence of acculturation.

Assessment and Disproportionality

To be in line with best practices, as well as legal and ethical requirements, those conducting psychological evaluations are concerned with the reliability and validity of the tests they use. Although reliability can readily be demonstrated during test construction, the validity of an assessment involves extraneous factors. Specifically, one would want to reduce or eliminate the likelihood that other factors are the reason for poor assessment results. This is precisely what the National Research Council (NRC; 1982) concluded when they stated the important of conducting assessments in two phases. The first phase should include a comprehensive assessment of the learning environment, which would include measures of whether a number of favorable instructional practices had been attempted. The second phase would also be detailed, but would be centered on the individual. Basically, the NRC's report supported the elimination of environmental or contextual factors prior to considering an individual assessment to determine eligibility for special education services. This process is of course challenging to practitioners as they are faced with an assessment procedure that is dependent on a number of factors. For example, "the number of students placed in special education will depend on the demands of the classroom, the resources to provide additional attention in the general education or compensatory education context, and the proclivities of teachers, administrators, and parents" (NRC, 2002, p.76).

Twenty years later, the NRC revisited the notion of equity/disproportionality in special and gifted education. In the executive summary of the more recent version, they questioned whether "there reason to believe that there is currently a higher incidence of special needs or giftedness among some racial/ethnic groups" (p.4). They go on to affirm this impression and provide support for their rationale from research on poverty and early development. In short, the panel contended that disproportionality that exists in special education services may not be the result of poor assessment practices, but rather may be strongly related to the high levels of poverty within culturally and linguistically diverse groups. Consequently, early childhood experiences of CLD children are more likely to involve exposure to harmful substances, poor nutrition and unstable home environments, all of which have a negative impact on early development (NRC, 2002). Conversely, some educators reacted to this comment as being an oversimplification of the issues related to CLD students and their disproportionate representation in special education. Specifically, it is the inner workings of the entire school system that affects students being over-represented in special education (O'Connor & Fernandez, 2006). This notion is supported, in part, by research demonstrating that there may be an association between the restrictiveness of the setting in which special education services are provided and the race/ethnicity of the student, regardless of disability category (Hosp & Reschly, 2004).

The role of teachers in the referral process and their resulting, albeit indirect, influence on students receiving special education services is not always considered because the focus of arguments is restricted to direct influences, such as test results.The frequency of teacher referrals may have a significant effect on special education placement, however, since a high number of students referred will be eligible for services (Ysseldyke, Vanderwood, & Shriner, 1997). Further, Harry and Klingner (2005) noted an interaction between teacher quality and referral frequency, with teacher quality as the underlying factor of concern. Teachers who use evidence-based instructional practices (e.g., differentiated instruction) and provide adequate support to students may only refer students whom they suspect are likely to have a disability due to their incapacity to respond to strong classroom instruction. In this case, a high likelihood of eligibility from a teacher referral would not be cause for concern. It is concerning, however, if teachers with high referral rates are not providing high quality instructional practices. In this situation it is difficult to rule out an insufficient opportunity to learn as the cause for low academic performance.

We can also see the interaction among instructional practices, teacher quality, and eligibility for special education services through questionable assessment results over time. *Brody v. Dare County Board of Education* exemplifies what could result from poor instructional practices—either from a general or special education teacher—and provides an example of a case built upon the results from a test of cognitive abilities. This case cites not only *The Matthew Effect* as grounds for parental concern, but also the subsequent court case that stemmed from a substantial decrease in IQ over time. James Brody's parents suspected that his decrease in IQ was due to insufficient educational practices that resulted in him being exposed to curriculum and academic rigor that was not comparable to his peers between his initial and subsequent re-evaluations. In this case, the attenuation of the reported IQ was attributed to a failure of a school to provide an adequate education to a student who was found to be eligible for special education services. His parents felt that James should have been receiving intensive remedial services that would help him demonstrate growth in academic areas identified through the evaluation process. *Brody v. Dare County Board of Education* is another example of consequential validity, but less of tests and more so of the process related to testing students for special education services. In the context of disproportionate representation of students in special education services, however, this different view of an attenuation over time, as opposed to an attenuation at the time of the initial referral, only further supports the notion of the complexity of the problem related to making high stakes decisions, such as determining eligibility for special education services.

It should be noted that the determination of eligibility in *Brody v. Dare County Board of Education* was based on a discrepancy between IQ and achievement (i.e. the discrepancy model), which has widely been disputed as a valid practice (Fletcher et al., 1998; Stuebing et al., 2002). Nonetheless, poor instructional practices appear to be at the core of this negative outcome that is demonstrated through cognitive testing. Unfortunately, this may be especially relevant to CLD students because schools that serve minority students with low socio-economic status tend to have the greatest extremes in the quality of instruction and classroom management (Harry & Klingner, 2005).

Disability Categories and Overrepresentation. The consequential validity of assessment practices continues to receive scrutiny as different trends in outcomes of evaluation procedures are observed. In recent years, there have been noteworthy changes in the trend of having African American males over-represented in being labeled with a disability and receiving special education services. At the national level, African Americans are still over-represented in the category of Emotional-Behavioral Disorder, but are *not* over-represented in the category of Learning Disability (Harry & Klingner, 2005). An explanation of this trend proposes that Learning Disabilities are considered an *elite* disability category, and that one of the underlying reasons for a lack of over-representation in this category is due to biases that Harry and Klingner (2005) consider a subtle form of segregation of African American students. Perhaps a more plausible claim would be to consider the incorporation of Response to Intervention (RtI) in schools and its influence on determining Learning Disabilities. Specifically, RtI reduces the reliance on the discrepancy model in making a determination of eligibility for special education services, but it does not necessarily apply directly to other disability categories, such as Emotional-Behavioral Disorders. Nonetheless, the discrepancy model could lead to inappropriate educational decisions if, for example, the IQ of an African American student was attenuated to the point where his or her score was too low to qualify him or her for services. Conversely, a score that is attenuated beyond the threshold of eligibility for a Developmental Cognitive Disability could lead to serious educational repercussions. For example, a student that may in fact have an IQ closer to the average range may receive services that are not at an appropriate educational level for at least three years or more, depending on the results of subsequent evaluations.

Hispanic Americans are not over-represented in *any* disability category at the national level (Harry & Klingner, 2005). This may be due to the increased awareness of English Language Learner status and a difficulty in ruling out the lack of English language knowledge in determining the existence of a disability. It may also be due to a lack of ability to demonstrate a discrepancy because scores of Hispanic Americans may be attenuated not only in areas requiring high verbal knowledge, but rather across all tests included in a cognitive battery due to an inability to fully understand directions and a lack of cultural understanding. Despite these national trends, some states and school districts continue to observe significant over-representation of minority groups receiving special education services.

Despite the controversy surrounding some of the statements made in their report, it appears that the NRC did identify the two major themes fueling the ongoing debate over disproportionality as invalid assessment procedures and poor instructional practices in the general and special education classroom (NRC, 1982). In short, there are a number of reasons why students may be performing poorly and identified as potentially having a disability that would make them eligible for special education services. However, before eligibility is determined, the aforementioned themes – invalid assessment and poor instructional practices – need to be addressed to ensure appropriate services are provided for struggling students.

The Assessment of Cognitive Abilities for CLD Students

Due to the ongoing increase in the number of CLD students enrolled in schools, we have seen a re-emergence of a heated debate that has been ongoing for over 30 years: the extent to which measures of intelligence are culturally or linguistically biased. The use of cognitive measures in school psychology practice and elsewhere is highly controversial, because bias in cognitive assessment may in fact be the most significant assessment issue to date (Brown, Reynolds, & Whitaker, 1999). Although Rhodes, Ochoa, and Ortiz (2005) present a number of recommendations to reduce the influence of Cultural Loading and Linguistic Demand during assessment procedures (), Jensen (1980) made a number of strong arguments against the existence of these influences. As we continue to move forward, it is clear that the selection and use of appropriate tests and the interpretation of the results are issues of great importance, especially when school psychologists use cognitive tests to make decisions about eligibility for special education services.

Historically, major concerns about bias in cognitive testing have been largely influenced by arguments that gene by environment interactions, as well as the technical *in*adequacy of cognitive measures, result in an underestimation of the true abilities of people from culturally and linguistically diverse backgrounds (Reynolds, 2000). These factors, combined with the unique influences of genetics and the environment, essentially represent the Cultural Test Bias Hypothesis (CTBH; Reynolds, 2000). Basically, this hypothesis contends that a number of factors contribute to a systematic underestimation of the true abilities of CLD individuals as measured by tests of cognitive abilities. Specifically, factors such as inappropriate content, inappropriate standardization samples, examiners' and language bias, and differential predictive validity were once thought to support this theory (Brown, Reynolds, & Whitaker, 1999). However, empirical evidence conducted in response to this hypothesis appears to reject the CTBH (Brown, Reynolds, & Whitaker, 1999). In particular, although there is some bias in cognitive measures, it actually works in favor of CLD individuals, as they actually score higher in some cases on items thought to be culturally loaded (Reynolds, 2000). Furthermore, expert panels that try to identify items that may be biased toward certain ethnic groups actually do a poor job of pinpointing items that in practice reduce test scores for those groups (Reynolds, 2000). Therefore, school psychologists should not be concerned with claims of bias toward CLD students when conducting assessments that incorporate the use of cognitive measures, as these measures have demonstrated superior technical adequacy and are robust to cultural bias (Brown, Reynolds, & Whitaker, 1999).

Clearly, the controversy is very heated and there are multiple perspectives to take into account. It becomes evident that we have yet to reach a consensus, as some remained unconvinced that cognitive measures are free of bias. It should be noted that although both sides take a strong stance on the issue, they are willing to admit that cognitive measures do have limitations that should be taken into consideration and that the debate, if it is to be resolved, should be clearly defined. For example, *situational bias* is often discussed with regard to cultural or linguistic bias, but it should not be used as an argument when debating the bias of cognitive measures because it represents a type of bias that is a product of the testing environment rather than the product of the test itself (Jensen, 1980). Nonetheless, the desire to reduce bias in any kind of testing is a valuable endeavor, particularly if the results of the assessment will have serious implications for the individual being tested. Therefore, as the debate over the existence of bias in cognitive testing continues among researchers, practitioners are left to grapple with judgments of appropriate measures and procedures when attempting to obtain the most reliable and valid measure of a student's abilities.

Current Practices in Assessing CLD Students

Many recommendations have been made to reduce the effects of cultural and linguistic loading of testing items. Completely eliminating bias is not possible, but practitioners should make it a goal to reduce it to the maximum extent possible. Professional associations have helped accomplish this goal by raising awareness of these issues in methods described in the *Standards for Educational and Psychological Testing* (American Educational Research Association, 2004) and the *Blueprint for School Psychology Practice* (Ysseldyke et al., 2006).

There are two interacting challenges when assessing CLD students – the effects of ethnicity/culture and the effects of language. When testing a student whose first language is not English, we need to be sure that "the assessment materials and procedures used actually assess students' target knowledge, skill, or ability, not their inability (or limited ability) to understand and use English" (Salvia & Ysseldyke, 2004, p.176). The effects of ethnicity are also highly variable, as there is a great deal of difference within any broad ethnic category. For example, two individuals who speak French, one from Canada and another from France, may be identified as both being Caucasian. Suppose they were both enrolled in an American school and given cognitive measures. They may experience differing levels of difficulty with the test because they not only have different dialects, one of which may be more in line with the translator's, but also they likely have varying levels of exposure to the American culture reflected in test items. These issues continue to influence the psychoeducational assessment of CLD students for special education services, as we continue to attempt to minimize their influence within our current practices.

Ideally, students should be tested in the language in which they are most proficient, which may or may not be their native language. If their native language happens to be the language in which they have the highest level of proficiency, there are a number of practical obstacles that reduce the likelihood of the test actually being administered in their native language. First, it is difficult to find a professional who speaks the native language of a student who is also adequately trained on a measure of cognitive abilities in that language. Second, the limited resources of schools typically do not make available to educational professionals multiple tests of cognitive abilities in a variety of languages. Third, there are only a limited number of cognitive measures demonstrating technical adequacy for educational decisions that exist in a number of languages.

It is not realistic to expect test developers to translate and norm their measures for every minority group in a given country. However, there are some tests that have been normed for Hispanic populations that have some experience with American culture. The extent to which a CLD student is able to adequately understand and respond to a cognitive measure in English is related to age, immersion in English, and the similarity of their native language to English (Salvia & Ysseldyke, 2004). Therefore, if a practitioner thinks that the student's experience with English relative to these three factors is limited, then they are likely to opt for an assessment that allows the student to use his or her native language at least to some extent. While practitioners are doing the best that they can to increase the validity of their assessment practices, they are still limited by the boundaries of their districts' resources. Consequently, many practitioners have opted for the translation of test directions and student responses, or have opted for nonverbal measures of cognitive ability.

Translation. There are significant concerns related to the practice of test translation or assessment in multiple languages (Rhodes, Ochoa, & Ortiz, 2005). Bias in this situation occurs three-fold. It can occur when the student's response is not accurately translated and induces increased measurement error in the scoring of his or her response. An extreme example of how this could occur is if the correct response to an item does not have a corresponding word in the subject's language. Bias can also occur when the translation alters the stimulus (i.e. the item of the test), which can change the difficulty of the item (Salvia & Ysseldyke, 2004) and, consequently, reduce the validity of the assessment. Finally, bias may occur if the content of the assessment may not be familiar to the student due to limited experience in the U.S. (Salvia & Ysseldyke, 2004). This likely comes into play when a student demonstrates a good level of language proficiency through reading and comprehension, which may cause the student's limited experience with the language to be overlooked. The result of any of these biases, which often results in a lower cognitive or achievement score, would then not reflect the student's true ability, but rather his or her inability to express him or herself in English. Therefore, in order to reduce this type of bias, school psychologists should take into account the effect that translating a test may have on the validity of its results.

Nonverbal Assessments. The use of nonverbal tests to adequately assess the abilities of individuals who do not possess high enough language abilities has been suggested as an appropriate procedure for reducing bias in tests heavily dependent on language (Bracken & Naglieri, in Reynolds & Kamphaus, 2003). This practice is questionable, however, as empirical evidence suggests that nonverbal tests are just as likely to be biased as verbal measures and also that verbal items may in fact perform better than non-verbal items in terms of technical adequacy (Reynolds, 2000). Furthermore, the tasks that typically require performance during a nonverbal measure of cognitive ability (i.e. the behaviors sampled by the test) tend to be less related to academic achievement than verbal cognitive measures (Salvia & Ysseldyke, 2004). This makes the procedure less relevant for the purposes of educational decision-making. Thus, we see why practitioners often struggle when making decisions about appropriate procedures when testing CLD students.

Although it can be difficult to address the multiple sets of issues related to appropriate assessment of CLD students, it is important that every student be tested fairly and that the school psychologist draws valid inferences from the results of the assessment. There are a number of ways in which practitioners can be sensitive to the biases that occur in testing a diverse group of students and engage in best practices when making educational decisions for these students. Ideally, a strong measure would substantially reduce measurement error and allow school psychologists to draw valid inferences from the results. Furthermore, it appears that a number of school psychologists do not feel comfortable with the interpretations that they draw when assessing CLD students, which they attribute primarily to a lack of training in this area (Rhodes et al., 2005). This finding should urge researchers to develop ways to reduce this effect by providing school psychologists with adequate tools to guide their interpretation of commonly used assessment procedures. Therefore, although many biases cannot be avoided in constructing and appropriately norming a test applicable across the country, an empirically-validated indicator of the extent to which each test is affected by cultural loading and linguistic demand would be extremely useful to practitioners as they attempt to make recommendations based on testing results.

The Culture-Language Interpretive Matrix (C-LIM)

Currently, the only broad guidelines available to practitioners to help them estimate the contribution of Cultural Loading and Linguistic Demand in influencing cognitive assessment results has been the Culture-Language Interpretive Matrix (C-LIM; Flanagan & Ortiz, 2001). It was originally conceptualized to answer the question of difference versus disorder from assessment results (Flanagan, Ortiz, & Alfonso, 2007). The difference versus disorder question is the distinction made in the attribution of results to differences in cultural and linguistic background or to an underlying disorder that is impairing cognitive functioning. The purpose of asking this question is to eliminate the potential confound of construct irrelevant variance. Construct irrelevant variance occurs when variance exists in scores that is not attributed to the construct the assessment method was intended to measure (e.g., cognitive abilities), but to an irrelevant source, such as race. The influence of construct irrelevant variance can be measured in a number of ways, such as analyses of differential item functioning. However, the C-LIM appears to go beyond broad generalizations when comparing one group to another in these types of analyses by attempting to account for individual differences in the extent to which language and culture affect performance on a given test from a cognitive battery.

Linguistic Demand. There is an obvious link between the ability of an individual to communicate in English and his or her ability to perform on a test given in English. For example, bilingual students tended to perform significantly lower than their monolingual peers on WISC-III tests that required verbal ability (Nieves-Brull, 2006).

Although this difference should not be attributed to psychometric bias, it does support the consideration on the Linguistic Demand of different tests within a cognitive battery and how this will affect scores. The Linguistic Demand of the test directions also vary a great deal in their complexity and length from one test to another for the WJ-III battery (Cormier, McGrew, & Evans, 2011). The Linguistic Demand of a test implies that the language demands of test items will interact with a number of factors such as (a) the age at which the test-taker learned English; (b) the linguistic background of the individual relative to those against which he or she is compared; and (c) the amount and quality of the exposure to English (Ochoa & Dynga, 2005). Construct irrelevant variance comes into play when, within a cognitive battery, tests "degenerate in unknown degrees into tests of English-language proficiency" (Figueroa, 1990, p.93).

Cultural Loading. The level of acculturation can be synonymous with culture, because it is likely to be the construct that will influence test scores the most. The familiarity with the culture (degree of acculturation) that is part of the test's norming group will likely have a positive impact on a person's test scores. This is one of the major assumptions of the C-LIM. Cultural bias is another important consideration that may come into play, regardless of the perceived linguistic bias of any given test. Often, the cultural bias that may occur is related to the level of acculturation experienced by the student being tested. The construct of cultural bias has best been described as more related to cultural loading in its effect on testing (Rhodes, Ochoa, & Ortiz, 2005). In practice, school psychologists may consider two factors to take into account the influence of cultural loading on certain test items: (a) a student’s level of acculturation compared to his or her peer group; and (b) the extent to which cultural knowledge influences performance on certain tests or items (Rhodes, Ochoa, & Ortiz). Some have suggested, however, that cultural loading of individual test items may in fact be a fallacy (Brown, Reynolds, & Whitaker, 1999). Unfortunately, it remains an important consideration as noted in the aforementioned professional standards and in discussions of negative effects of testing CLD students. Therefore, at the very least, a consensus could be reached that more research is needed to either support or refute the notion of cultural bias in testing.

Those describing the C-LIM in detail (Flanagan, Ortiz, & Alfonso, 2007) outline the degree of attenuation that is expected from three groups of CLD individuals. A distinction is made among those *slightly different*, *moderately different*, and *markedly different* (see Flanagan et al., 2007, p.199). The important contributors used in determining the extent to which a student is different from the general U.S. population are the level of English proficiency and the amount of time in the country, which relates to his or her degree of acculturation. Students’ scores become more attenuated in relation to the extent to which they are markedly different on one or both of the dimensions of Cultural Loading or Linguistic Demand. The degree of attenuation suggested by the developers of the C-LIM ranges from 3 to 5 standard score points for an individual identified as *low* on both dimensions to an attenuation of 25 to 35 standard score points for those identified as high on both Cultural Loading and Linguistic Demand. Figure 1 shows this attenuating effect, which can be applied to the tests categorized in Figure 2.

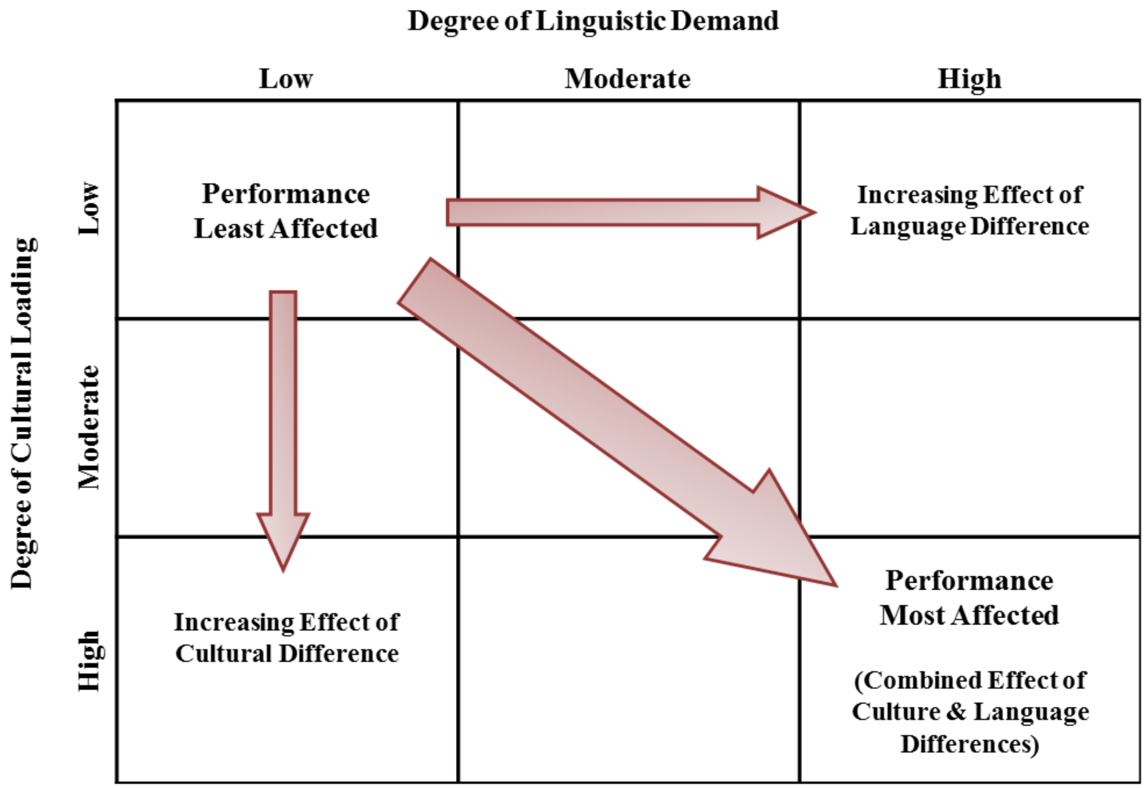

Figure 1. Pattern of predicted performance for CLD students based on C-LIM classifications (Flanagan, Ortiz, & Alfonso, 2007).

Unfortunately, the classification of the tests presented in the C-LIM (see figure 1) was mostly based on expert judgment, and only limited empirical evidence currently exists to offer partial support for the classification of tests from a number of common test batteries. However, preliminary evidence examining the Linguistic Demand component of the C-LIM suggests that Linguistic Demand may be more accurately represented as a multidimensional construct involving measures of receptive and expressive language (Hansen, Cormier, Lim, & McGrew, 2010). Although the influence of culture on test scores has been heavily researched (Reynolds, 2000; Jensen, 1980), researchers have yet to model the potential effects of culture and language simultaneously. This is a critical unmet need considering "language and culture are inextricably linked" (Ortiz & Dynda, 2005, p.549).

Linguistic Demand → / Cultural Loading ↓	Low	Moderate	High
Low	Spatial Relations	Visual Matching Numbers Reversed	Concept Formation Analysis Synthesis Auditory Working Memory
Moderate	Picture Recognition Planning Pair Cancellation	Visual Auditory Learning Delayed Recall-Visual Auditory Learning Retrieval Fluency Rapid Picture Naming	Memory for Words Incomplete Words Sound Blending Auditory Attention Decision Speed
High			Verbal Comprehension General Information

Figure 2. Cultural-Linguistic Interpretive Matrix for the Woodcock-Johnson Test of Cognitive Abilities (Rhodes, Ochoa, & Ortiz, 2005).

Preliminary Empirical Evidence to Support C-LIM Classifications

Currently, only one empirical study of the C-LIM has been published in a peer-reviewed journal. In addition, there are a few unpublished dissertations that also provide some preliminary evidence on the classifications of the C-LIM. Previously completed research attempted to validate the C-LIM classifications by demonstrating the predicted downward trend in test performance for CLD students, when compared to other students with or without disabilities. Verderosa (2007) examined the extent to which test performance of CLD individuals on the Differential Abilities Scale (DAS) reflected the C-LIM classifications. In this study, Verderosa (2007) compared the performances of preschoolers with differing levels of English language proficiency (e.g. Spanish only, bilingual, English only) on the DAS. The observed differences in test performance appeared to provide support for the pattern of predicted performance seen in Figure 1. It should be noted, however, that Verdorosa's (2007) sample was relatively small. Further, Verdorosa did not adjust the p-values to determine the significance of the differences between groups, which is likely to produce a number of Type II errors.

Tychanska (2009) obtained an archival dataset containing test scores of CLD students to make comparisons similar to Verderosa (2007) using the Wechsler scales, which includes the Wechsler Intelligence Scale for Children, Third Edition (WISC-III), Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV), and Wechsler Preschool and Primary Scale of Intelligence, Third Edition (WPPSI-III). Tychanska (2009) ran a large number of *t*-tests \to compare test performance of different groups against individual tests to the predicted (mean) performance. Tychanska (2009) observed the C-LIM's expected declining pattern of performance when comparing native English speakers and English Learners with and without learning or speech disabilities. However, , Tychanska (2009) also did not adjust p-values to determine statistical significance of the differences between groups.

Although these two dissertations provided some preliminary evidence to support the C-LIM classifications, Kranzler, Flores, and Coady (2011) observed the expected downward trend in only 13% of their sample of CLD students. Further, most of their analyses suggested that the differences in test scores were not significant, which lead them to conclude that further evidence to support the matrix classification system is needed to support its practical use. The lack of the expected trend in score patterns suggested by the C-LIM may not be due to an absence of Cultural Loading and Linguistic Demand occurring, but rather to a misclassification of the tests. Basically, if the classification of the tests in the C-LIM is inaccurate, this would lead to incorrect cell values being calculated in attempting to observe the downward trend. As a consequence, the classification of the C-LIM should be evaluated empirically before further research is conducted to validate its classification. Further, there has yet to be a study of the C-LIM classification of the WJ-III battery that incorporates both Cultural Loading and Linguistic Demand.

Building upon the literature and the current need in research and practice to better understand and address issues related to cultural loading and linguistic bias in cognitive testing, the purpose of this study is to:

- Determine the extent to which culture and language ability influence scores on each of the individual tests that measure cognitive abilities.

- Extend previous research by examining the extent to which receptive and expressive language abilities separately contribute to differences in individual tests scores.
- Use the information to provide a comparison between tests to determine the extent to which the categorization in the C-LIM is appropriate for use by researchers and practitioners.

Chapter 2 – Methods

Sample

The Woodcock-Muñoz Foundation provided the data from the WJ-III NU norming sample for use in this project. The norming sample included 8,782 subjects who were recruited from more than 100 geographically diverse U.S. communities. The original sample consisted of 1,153 preschool children (ages 2-5 and not enrolled in kindergarten), 4,740 students enrolled in K-12, and 2,889 adults (1,727 who were not enrolled in college or university and 1,162 undergraduate and graduate students). All subjects took the Woodcock-Johnson Test of Cognitive Abilities, third edition, and the Woodcock-Johnson Tests of Achievement for the purposes of co-norming the two tests. This study used a subset of the complete sample, which included students from ages 7 to 18. Table 1 shows the sample sizes according to age group and race.

Table 1

Sample Sizes by Race Category and Age Group

	Age Groups		
Race	7-10	11-14	15-18
White	1,261	922	796
Black	269	196	177
Indian[1]	45	28	18
Asian/Pacific Islander[2]	98	76	48
Hispanic	202	173	101
Total	1,875	1,389	1,140

[1]Includes American Indian, Eskimo, Aleut

[2]Includes Chinese, Filipino, Japanese, Asian Indian, Korean, Vietnamese, Other Asian, Pacific Islander

Measures

Woodcock-Johnson Test of Achievement, Third Edition (WJ-III ACH). The WJ-III ACH is a norm-referenced individually-administered battery of tests aimed at measuring overall academic abilities and oral language (McGrew, Schrank, & Woodcock, 2007). The WJ-III ACH provides not only a score related to total achievement, but also provides subject-specific scores in mathematics, written language, oral language, and other clusters such as academic skills, academic fluency and academic knowledge. Some of the narrow CHC abilities that are tested through achievement areas are (a) reading speed; (b) listening ability; (c) numerical facility; (d) spelling ability; and (e) writing ability. Further, content-specific test requirements include abilities such as identifying printed letters and words, adding, subtracting, multiplying rapidly, formulating and writing simple sentences rapidly, and identifying a missing keyword that makes sense in the context of a written passage (McGrew, Schrank, & Woodcock, 2007). The stimuli presented to the individual being tested can be visual or auditory and can be presented in text, question, or numerical form.

The reliabilities of cluster scores for academic content areas across all ages within the norming sample ranges from .93 to .97, which are within the recommended range for making significant educational decisions. The individual tests used in this study (*Understanding Directions, Oral Comprehension; Story Recall, and Picture Vocabulary*) have reliabilities ranging from .83 to .87. The validity evidence for this measure is extensive and the technical manual provides information on construct and content coverage for each of the individual tests (McGrew, Schrank, & Woodcock, 2007).

Woodcock-Johnson Test of Cognitive Abilities, Third Edition (WJ-III COG). The WJ-III COG is a norm-referenced tool used to measure overall cognitive abilities, as well as verbal and non-verbal abilities. The WJ-III COG assesses areas such as verbal comprehension, perceptual reasoning and working memory. Verbal comprehension is the ability to understand and create messages. Examples include familiarity with sentence structures, word problems and vocabulary. Perceptual reasoning is the ability to take in visual-spatial information to solve problems. An example is the ability to rotate a shape to fit it into a puzzle. Working memory is a brain function in which information can be retained temporarily as it is being formed, transformed, or executed. Complex cognitive tasks that involve learning, reasoning, and comprehension use working memory. An example is being told a forward sequence of numbers and having to repeat the numbers in reverse order.

The reliability coefficients of the clusters from the standard battery of the WJ-COG range from .90 to .97. The reliability for individual tests rangefrom .81 to .97. As was the case with the WJ-ACH, the validity evidence for tests included in the WJ-COG standard and extended battery is comprehensive and described in detail in the technical manual (McGrew, Schrank, & Woodcock, 2007).

Woodcock-Johnson Norming Update. Demographic information was also used from the Woodcock-Johnson Norming Update sample. Researchers aimed to include a representative sample of the U.S. population aged 12 months to over eighty years of age. They selected subjects using a stratified sampling design with random selection that controlled for the following: census region, community size, sex, race, type of school, type of college/university, education of adults, occupational status of adults, foreign born, and occupation of adults in the work force.

The sampling procedure occurred in three stages. The first stage included the sampling of communities that were representative of the U.S. population distribution by census region. The first stage of the sampling procedure also took urbanization and socio-economic status into account. Stage 2 included the sampling of available pools of students that were representative of a cross section of the entire community. The sample also included parochial and private schools. The third and final stage involved the random selection of school-aged subjects based on quotas set per grade level. The sample included students with disabilities as long as they received at least part-time instruction in the general education classroom. The sample excluded students who had less than one year's experience in regular English-speaking classrooms, however. Homeschooled students participated in the norming sample. The tests developers collected the data for school-aged subjects from September, 1996 to May 1999, for college/university students from September 1996 to March 1999, and for preschool and adult subjects from September 1996 to August 1999. The test developers then subjected the sample to a normative update which involved a "recalculation of the normative data, based on 2005 U.S. Census Statistics" (p. 1, McGrew, Schrank, & Woodcock, 2007).

Variables

Exogenous Variables. In latent variable structural equation modeling (LVSEM), exogenous variables only have paths going away from them and have no incoming causal paths. In this case, the first phase of the analysis included the exogenous variables Cultural Loading and Linguistic Demand. The latent variable for Linguistic Demand included four measures from the WJ-III Tests of Achievement: (a) Understanding Directions; (b) Oral Comprehension; (c) Story Recall; and (d) Picture Vocabulary. The measure of Cultural Loading included a four-factor model including: (a) Foreign Born Status; (b) Race; (c) Language at Home; and (d) First Language. The researcher selected these four variables because they are either innate or occur early in life and the shared variance between them is likely to capture, at least to some extent, the culture of a given individual. The second phase also used the variable of Cultural Loading, with the same variables used in Phase 1. However, the researcher split the latent exogenous variable for Cultural Loading into two distinct, exogenous variables – Expressive Language and Receptive Language.

Endogenous Variables. Endogenous variables either have causal paths coming to and going away from them, or only have incoming causal paths. In other words, endogenous variables can be both independent and dependent variables. Both phases used the standard scores from the WJ-III NU as an endogenous variable. The researcher fit and compared models with each of the following individual tests being evaluated: Spatial Relations, Picture Recognition, Planning, Pair Cancellation, Visual Matching, Numbers

Reversed, Visual Auditory Learning, Delayed Recall-Visual Auditory Learning, Retrieval Fluency, Rapid Picture Naming, Concept Formation, Analysis Synthesis, Auditory Working Memory, Memory for Words, Incomplete Words, Sound Blending, Auditory Attention, Decision Speed, Verbal Comprehension, and General Information.

The researcher created Expressive and Receptive Language variables in Phase 2 based on the same four tests for Linguistic Demand, with the latent variable Expressive Language composed of the WJ-III ACH tests of: (a) Story Recall; and (b) Picture Vocabulary. The remaining two tests from the Linguistic Demand variable in Phase 1, Understanding Directions and Oral Comprehension, were factors for the latent variable Receptive Language.

Procedure

The researcher used latent variable structural equation modeling to fit the general model outlined in Figure 2. The researcher evaluated the latent variables and paths within this model using the AMOS program to determine the overall fit and the strength of their respective path coefficients. This approach was selected due to the common use of AMOS and other structural programs for modeling relationship among latent variables (Van der Vijver & Poortinga, 1997). Furthermore, this method allowed for a simultaneous test of the influence of Cultural Loading and Linguistic Demand on scores for tests of cognitive abilities.

It should be noted that due to the sampling method used in the norming of the Woodcock-Johnson, not all variables were available for all subjects. Therefore, variable sample sizes varied across different individual tests. In addition, it was necessary to estimate means and intercepts for the model due to the missing cases. The researcher considered a multiple imputation with a subset of the available data. The results of the imputation were very similar to those observed with estimating means and intercepts. Consequently, the researcher opted for the estimation method to simplify data analysis and reduce the likelihood of problems incurred from additional steps required for the imputation.

The use of the function to estimate means and intercepts required the use of Maximum Likelihood Estimation (MLE) to generate the parameters for the model. MLE assumes that variables in the model are normally distributed. This did not prove to be problematic for the current analyses, because the sampling method used and the underlying assumptions that cognitive and academic abilities are normally distributed in the population. Consequently, MLE provided an estimation of the parameters that would be observed for the population from the representative Woodcock-Johnson norming sample data.

Age Groups. The way in which children and adolescents interact with the items presented during assessment procedures changes as they age in response to cognitive development and learning. Thus, in order to create an accurate model of the relationships between culture, language, and test scores, the researcher organized the data to reflect developmental stages. Specifically, the data were categorized into age groups according to Piaget’s Theory of Cognitive Development (see Oakley, 2004, for more information) .

According to this theory, the concrete operational stage occurs from ages 7 to 11 and is characterized by thinking that is more logical, objective, and deductive than in the previous stage, but which is not yet abstract. The Formal Operational stage is characterized by abstract thinking and is typically observed in individuals 11 or 12 years and up. Arlin (1975) proposed a fifth stage during the young adult years, which divides the Formal Operational stage into two distinct stages – the problem-solving stage (the Formal Operational stage) and the problem-finding stage. The benefit of using Piaget's theory is that the cognitive structures that he describes are "content independent and also domain general" (Wellman & Gelman, 1992, p.339). As a result, the theory can be applied to a multitude of areas influenced by cognitive development, such as the assessment of cognitive abilities.

Piaget's Theory of Cognitive Development also coincides well with academic milestones: The concrete operational stage aligns with typically developing students in the early elementary years (i.e. ages 7 to 10 being in grades 1-4); the formal operational stage aligns with student cognitive development in the late elementary and middle school years (i.e. ages 11 to 14 in grades 5-8); and further cognitive development, as suggested by Arlin (1975), is associated with cognitive functioning of students in high school (ages 15 to 18 in grades 9-12). There are significant shifts in the curricular expectations of students as they move from one milestone to another. Thus, these categories not only attempt to capture stages of cognitive development, but also account for the learned behaviors that result from increasing curricular expectations across the academic milestones.

Cross-validation. The researcher split the data randomly into 2 separate files. This split allowed for post-hoc model fitting. The researcher conducted a cross-validation following post-hoc fitting, as needed, to provide validity evidence for the results observed with the test sample. The split was completed using the *Select Cases* function in SPSS version 19.0. This function takes a random set of cases of approximately half the data set and saves them to a separate file. The researcher completed the split procedure for the three aforementioned age groups, which resulted in six files being used for data analysis. Figure 3 and Figure 4 outline the structures of the multiple datasets.

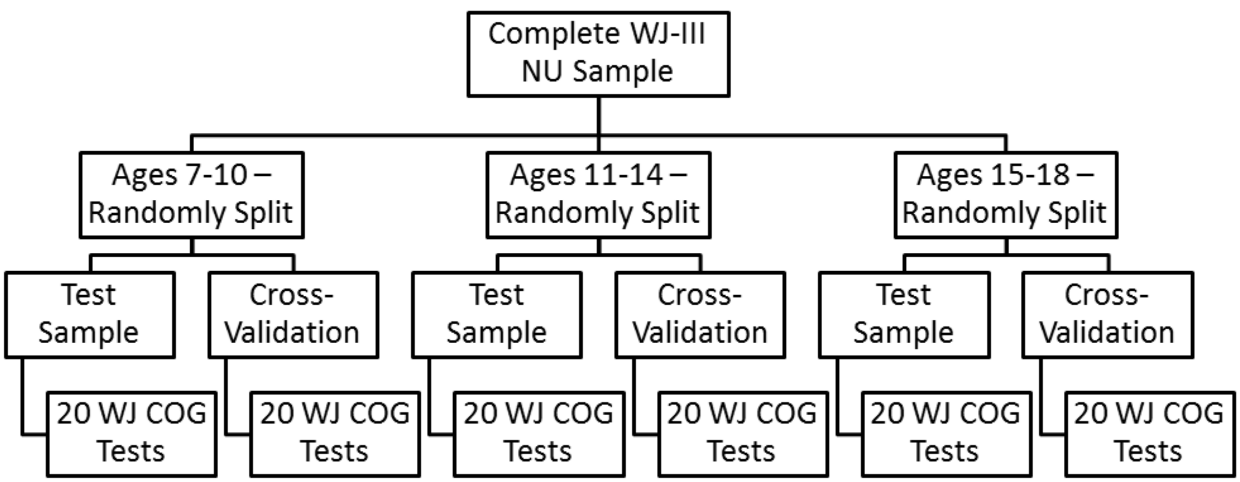

Figure 3. Data analysis organization for Phase 1.

In summary, the structure for Phase 1 (see Figure 1) allowed for a comprehensive evaluation of the influence of Cultural Loading and Linguistic Demand on each of the 20 individual tests that are included in the standard and extended battery of the WJ-III COG. Further, as mentioned previously, the structure accounted for different levels of cognitive development. In Phase 1, the individual datasets for the 3 age groups (e.g. 7 to 10, 11 to 14, and 15 to 18) were randomly split into test and cross-validation samples. Then, the researcher modeled each of the 20 individual tests within each of these samples. Thus, the total number of models run for Phase 1 was determined by adding each cluster of 20 individual tests at the bottom of the structure presented in Figure 1, resulting in 120 models run for Phase 1. Phase 2 had a similar structure, although the total number of models was doubled due to the influence of expressive and receptive language being tested individually. Consequently, the researcher ran 240 models in Phase 2 of the study (see Figure 2).

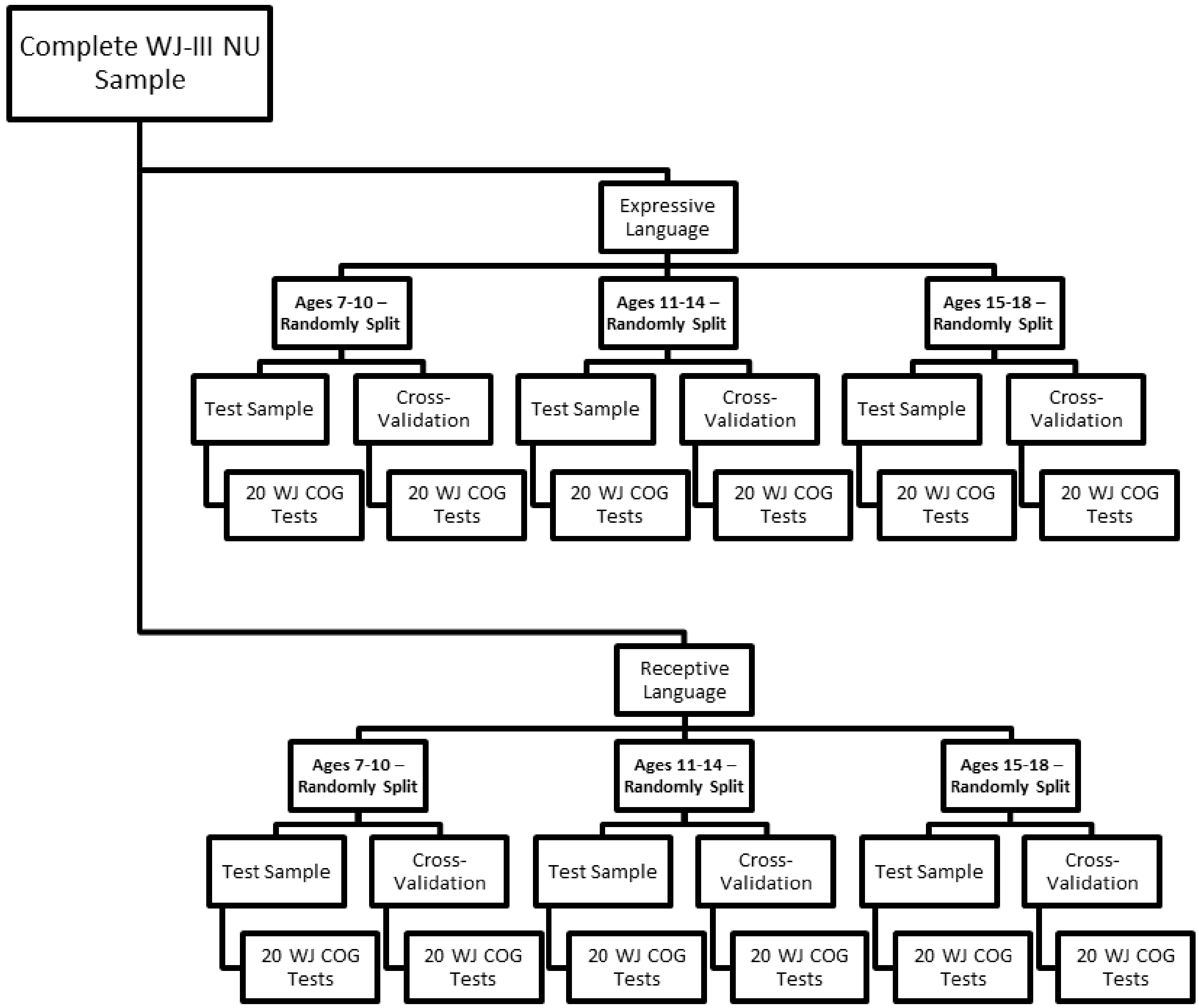

Figure 4. Data analysis organization for Phase 2.

Coding. Some of the variables in the dataset needed to be re-coded for the analysis. The researcher recoded three measures for the latent variable Cultural Loading (Foreign Born Status, Language at Home, and First Language) to be dichotomous, primarily because of the low sample sizes of their original categorical coding. Foreign born status received a dummy code, with the value of 1 representing non-U.S. born and 0 representing U.S. born. Language at Home and First Language also received dummy codes with 1 representing Not English and 0 representing English. The remaining variable for the Cultural Loading latent variable, Race, used a categorical coding. All other variables in the measurement model maintained their original coding (see Table 2).

Table 2

Summary of Scale and Coding of Key Variables

Latent Variable and Indicator	Scale/Format	Coding
Cultural Loading		
Foreign Born Status	Dichotomous	0 = U.S. born 1 = Non-U.S. born
Race	Categorical	1 = White 2 = Black 3 = Indian 4 = Asian/Pacific Islander 5 = Hispanic
Language at Home	Dichotomous	0 = English 1 = Not English
First Language	Dichotomous	0 = English 1 = Not English
Linguistic Demand		
Story Recall	Scale	Not Applicable
Picture Vocabulary	Scale	Not Applicable
Understanding Directions	Scale	Not Applicable
Oral Comprehension	Scale	Not Applicable
Receptive Language		
Story Recall	Scale	Not Applicable
Picture Vocabulary	Scale	Not Applicable
Expressive Language		
Understanding Directions	Scale	Not Applicable
Oral Comprehension	Scale	Not Applicable
Individual Test Score		
20 Individual Tests	Scale	Not Applicable

Phase 1 Analysis

Figure 3 shows the generic model tested in this phase. The model is a representation of the theoretical influence of cultural loading and Linguistic Demand on test scores from the Woodcock-Johnson Test of Cognitive Abilities that is thought to exist based on the Culture-Language Interpretative Matrix (Rhodes, Ochoa, & Ortiz, 2005). The C-LIM model suggests a simultaneous contribution of both cultural loading and Linguistic Demand on total test score. Further, the two axes of the matrix suggest an interaction between Cultural Loading and Linguistic Demand, as they both have varying degrees of influence depending on the test. Consequently, in order to appropriately represent this interaction, the researcher allowed the latent variables Cultural Loading and Linguistic Demand to co-vary. Each test from the WJ-III COG battery was tested within its own model. After the researcher had completed initial model testing, he did some post-hoc fitting to address any issues related to co-linearity. After this step, the researcher ran a cross-validation on the final model.

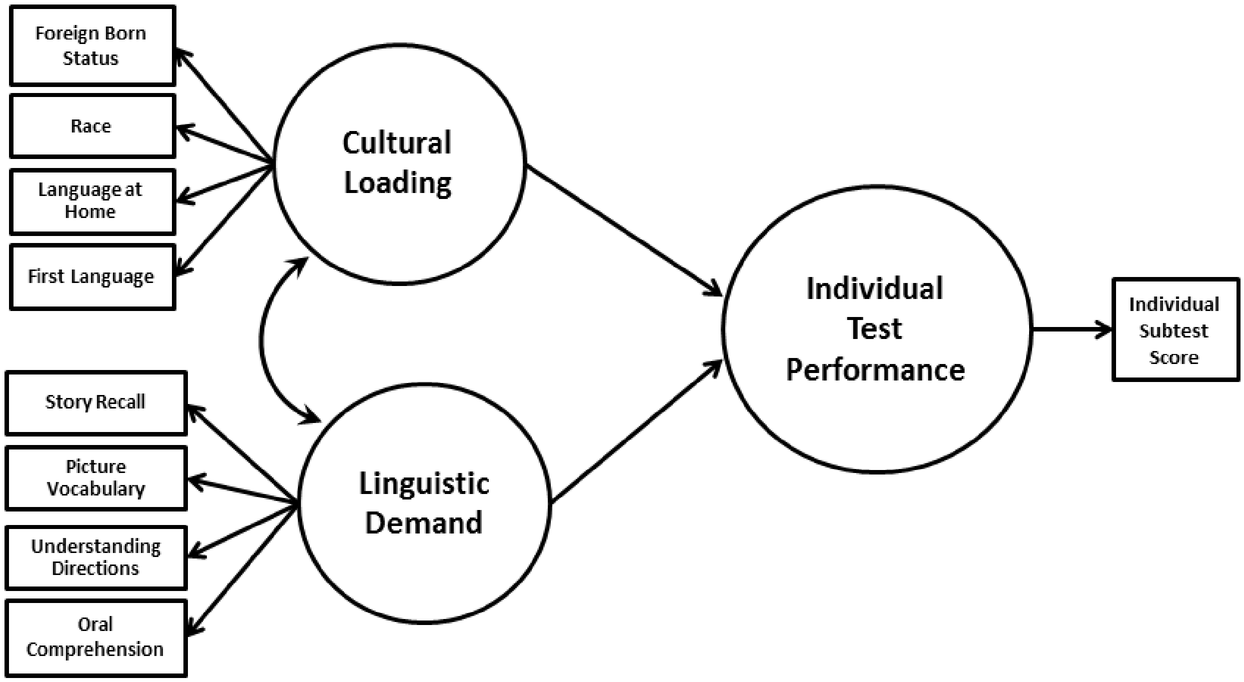

Figure 5. Hypothesized general model representing the relationships between cultural loading and linguistic demand with individual test scores.

Phase 2 Analysis

The purpose of Phase 2 was to build upon the information provided by Phase 1 and suggest a revised model that would possibly better represent the relationship between the influences of Cultural Loading, Linguistic Demand, and test scores. Figure 4 shows the general model that was originally developed for this phase. The purpose of the structure of the revised model was to better represent the relationship between language and individual test scores by modeling expressive and receptive language independently. Unfortunately, preliminary analyses revealed that multi-colinearity was preventing the software from solving this initial model and providing parameter estimates. Thus, the researcher separated the procedure into two sets of analyses – one that would only model the influence of receptive language and another that would only model the influence of expressive language.

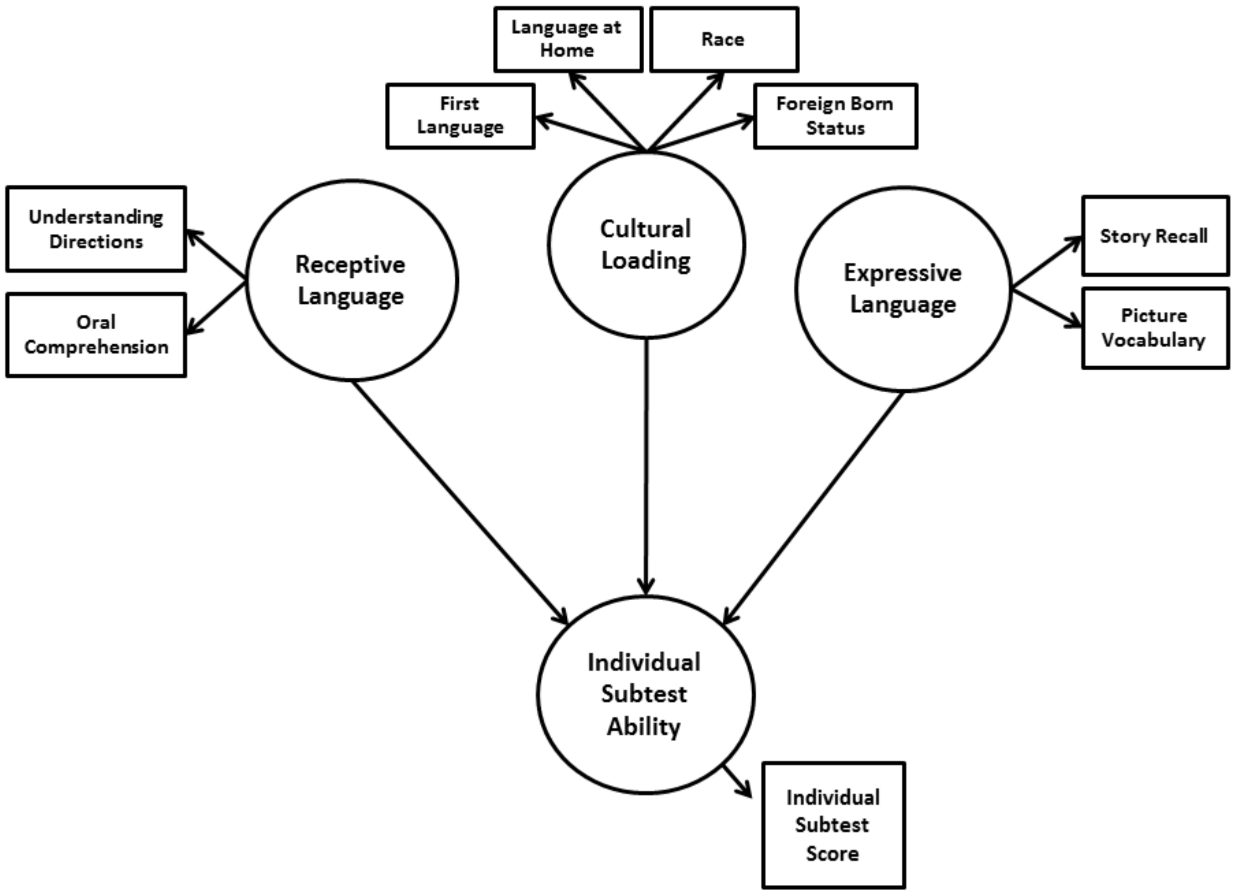

Figure 6. Original model for Phase 2 with individual latent variables for expressive and receptive language.

Figure 5 shows the revised models run for this phase. Another revision made subsequent to the original model was the usage of the Cultural Loading variable as an exogenous variable with only a direct causal path to the language variable (either expressive or receptive language). This was due to the relatively low coefficients for the direct path observed in the models run for Phase 1. Further, because of the empirical evidence suggesting relatively low influence of cultural bias in test items (Jensen, 1980) and the indirect nature of cultural influence, it may be reasonable to consider Cultural Loading as having an indirect influence on cognitive test scores. This differs from the theoretical model suggested by the C-LIM. However, given the parameters in Phase 1, it appears that Linguistic Demand should be weighed more heavily than Cultural Loading, which is contrary to the C-LIM model. Nonetheless, the researcher considered culture as an indirect influence, and this model's purpose was to provide empirical evidence to further support the results from Phase 1 in weighing linguistic ability more heavily than cultural loading when making decisions about test selection.

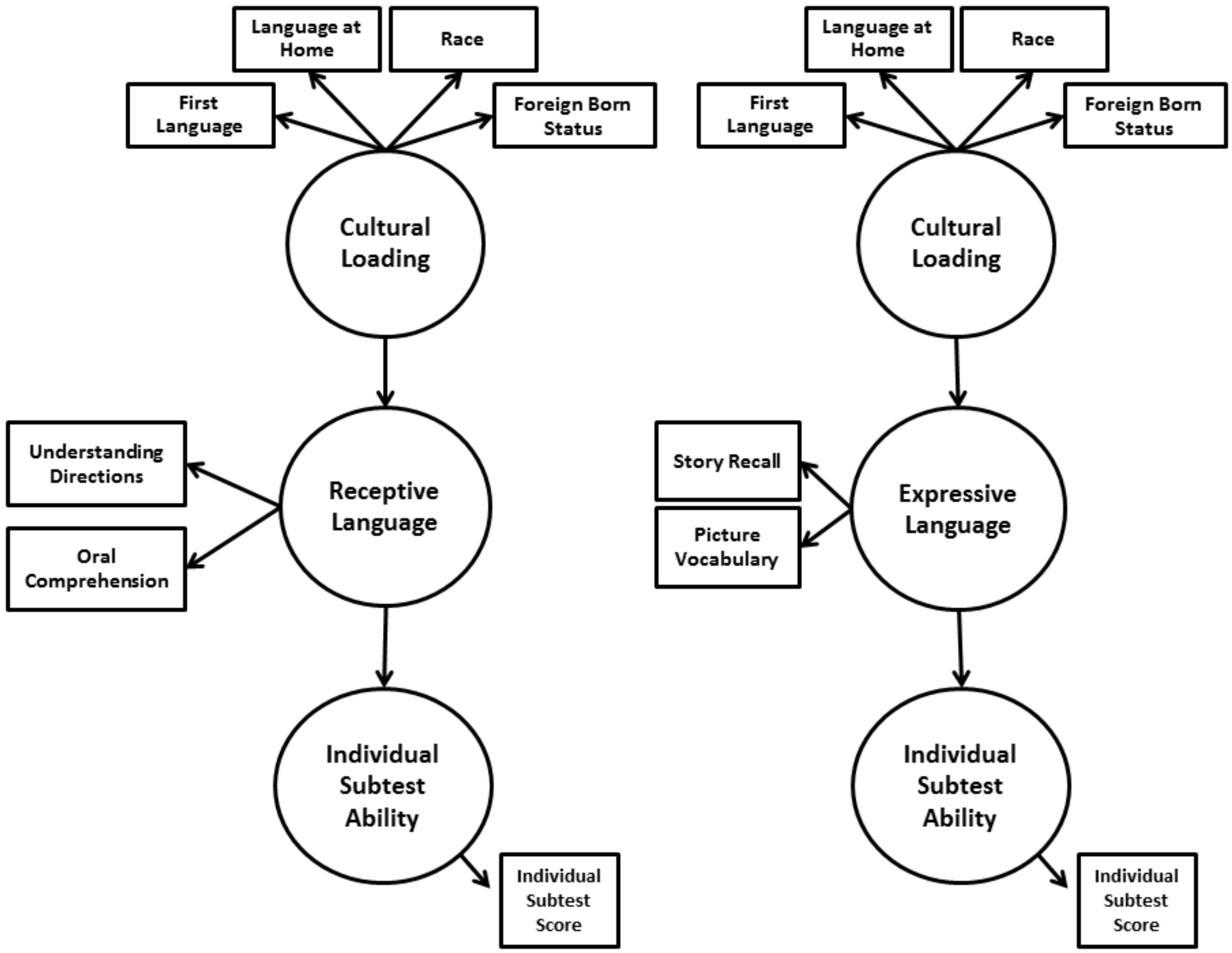

Figure 7. Revised Phase 2 models evaluating the influence of expressive and receptive language individually.

Chapter 3 – Results

Overview

Across both phases of this study, the researcher used latent variable structural equation modeling (LVSEM) to represent the theoretical relationships among Cultural Loading, Linguistic Demand, and performance on individual tests from the Woodcock-Johnson Test of Cognitive Abilities battery according to the *Culture-Language Interpretative Matrix*. The study consisted of two phases – one that examined the relationships between Cultural Loading and Linguistic Demand on individual test performance, and another that examined the unique contribution of expressive and receptive language on individual test performance. The purpose of Phase 1 was to determine the extent to which Cultural Loading and Linguistic Demand affect test scores on a measure of cognitive abilities. The purpose of Phase 2 was to determine the extent to which expressive and receptive language abilities contribute to performance on individual tests. Overall, the results will yield empirical evidence to provide guidance to practitioners in the varying use of individual tests from the WJ-III battery based on the cultural and linguistic background of the student being assessed. Due to the high level of consistency between initial model tests and the cross-validation sample, only the results from the cross-validation are shown and discussed in this section. However, all of the data from the entire data analysis can be found in Appendices A, B, C, and D.

Phase 1

Model Fit. The researcher evaluated model fit to determine how consistent the observed data were with the proposed model. In this case, the data appeared to fit relatively well across tests and age groups. The researcher obtained a rough estimate of how well the data fit the model by dividing the Chi-Square value by the degrees of freedom for the model. A good model fit is established if the results of the rough estimate are less than 4. In general, the values of this calculation across tests and age groups ranged from around 4 to 6. However, due to the influence of sample size on calculating Chi-Square statistics, the researcher gathered additional fit information from a number of fit indices such as the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), and the Normed Fit Index (NFI). The RMSEA evaluated fit between non-nested models with a good fit valueof.10 or less. The CFI and NFI evaluated fit between nested models with a good fit value of .90 or more. Based on these three indices, it appears that the data fit these models well across individual tests and age groups. For information on specific fit indices, see Appendix A.

Path Coefficients. Results were relatively consistent across age groups. The path coefficients from Linguistic Demand to Individual Test Performance demonstrated a relatively large range across individual tests for all three age group models. The path coefficients from Cultural Loading to Individual Test Performance, however, had a low value and were relatively restricted in their range. In LVSEM, a path coefficient that is less than .30 is considered to not demonstrate a significant relationship between variables.

Thus, the path coefficients between the latent variables Cultural Loading and Individual Test Performance were not considered to be significant. In addition, although there were relatively consistent covariance coefficients between the latent variables Cultural Loading and Linguistic Demand, these coefficients seldom exceeded .30, and the maximum value for these coefficients across all models was .37.

As seen in Figure 8, the individual tests that had relatively high values for the path coming from the latent variable Linguistic Demand were General Information ($r = .79$), Concept Formation ($r = .82$), and Verbal Comprehension ($r = .90$). The individual tests that had relatively low path coefficients or values that did not meet the threshold for significance (i.e. path coefficient less than .30) were Picture Recall ($r = .14$), Planning ($r = .26$), and Spatial Relations ($r = .28$).

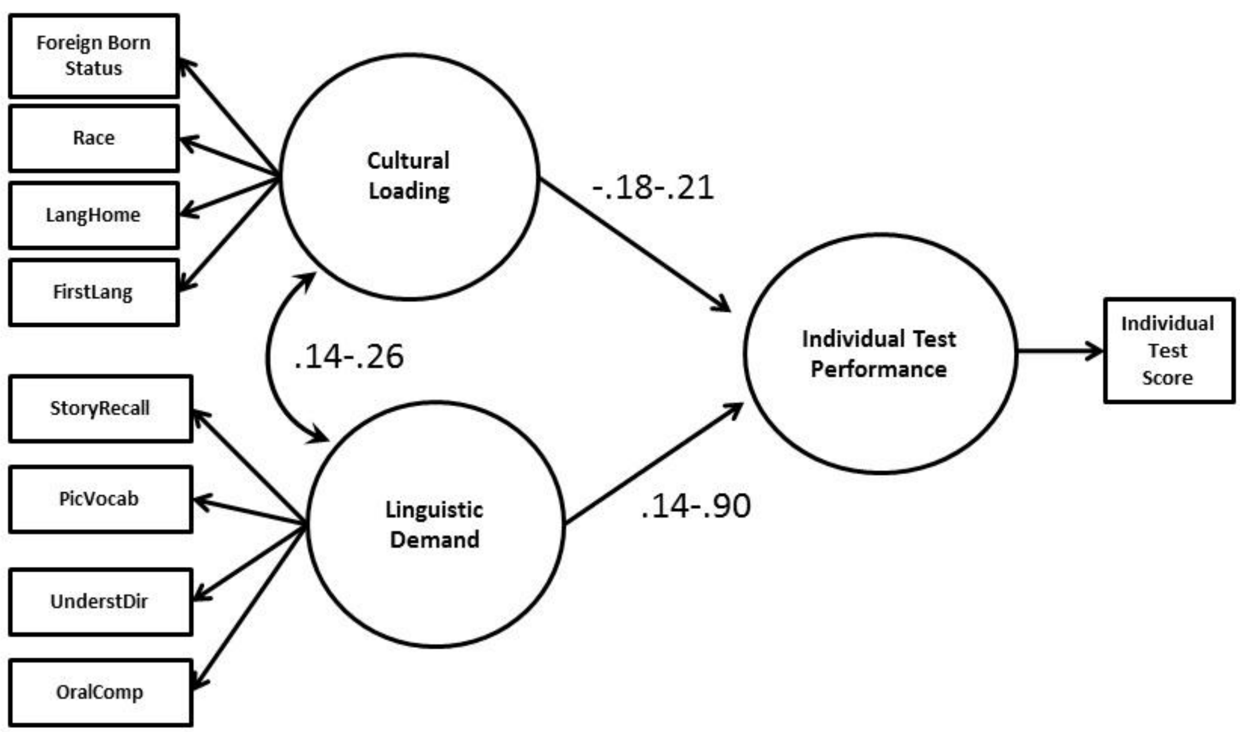

Figure 8. Path coefficient ranges from latent variable structural equation models for Phase 1, ages 7 to 10.

As seen in Figure 9, the 11 to 14 age group exhibited similar results. The individual tests that had relatively high values for the path coming from the latent variable Linguistic Demand were Concept Formation (*r* = *.86*), General Information (*r* = *.91)*, and Verbal Comprehension (*r* = *.94*). The individual tests that had relatively low path coefficients or values that did not meet the threshold for significance (i.e. path coefficient less than .30) were Rapid Picture Naming (*r* = *.25*) and Picture Recall (*r* = *.26*). Appendix B provides all the results for the individual path coefficients across individual tests and across age groups.

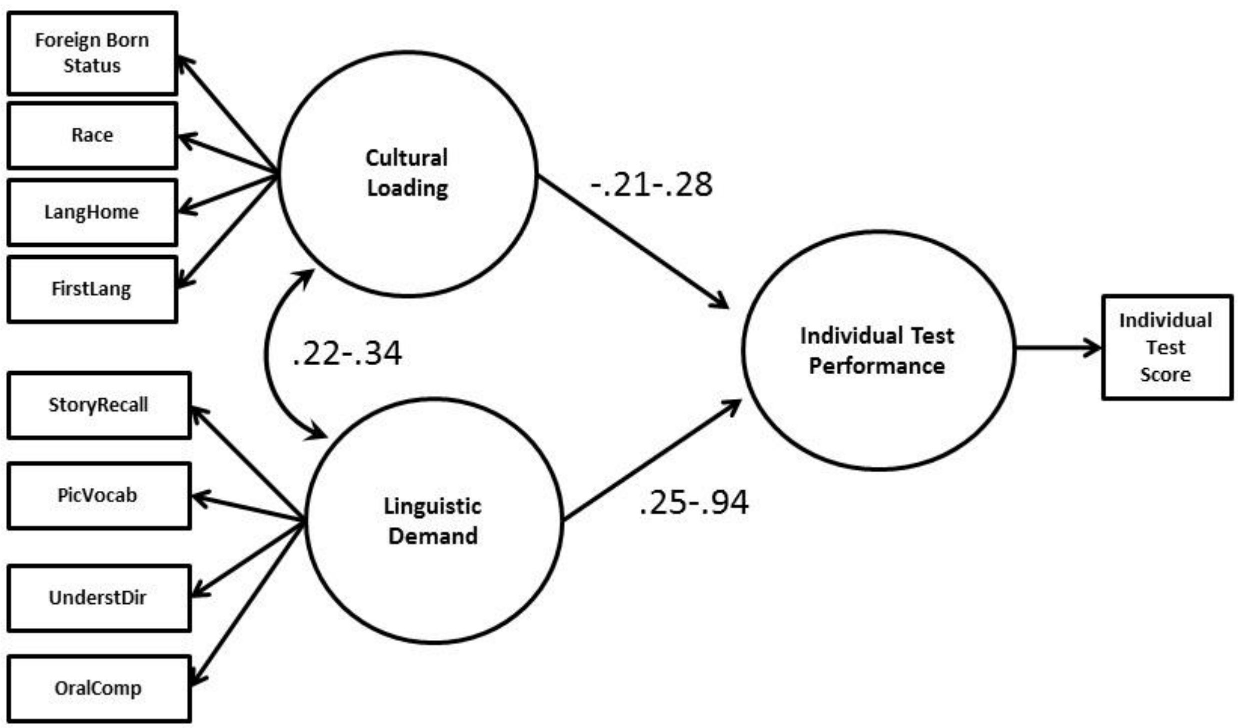

Figure 9. Path coefficient ranges from latent variable structural equation models for Phase 1, ages 11 to 14.

As seen in Figure 10, the 15 to 18 age group showed similar results. The individual tests that had relatively high values for the path coming from the latent variable Linguistic Demand were Concept Formation ($r = .83$), General Information ($r = .90$), and Verbal Comprehension ($r = .90$). The individual tests that had relatively low path coefficients or values that did not meet the threshold for significance (i.e. path coefficient less than .30) were Auditory Attention ($r = .24$), Pair Cancellation ($r = .32$), and Picture Recall ($r = .33$). Appendix B contains all the results for the individual path coefficients across individual tests and across age groups.

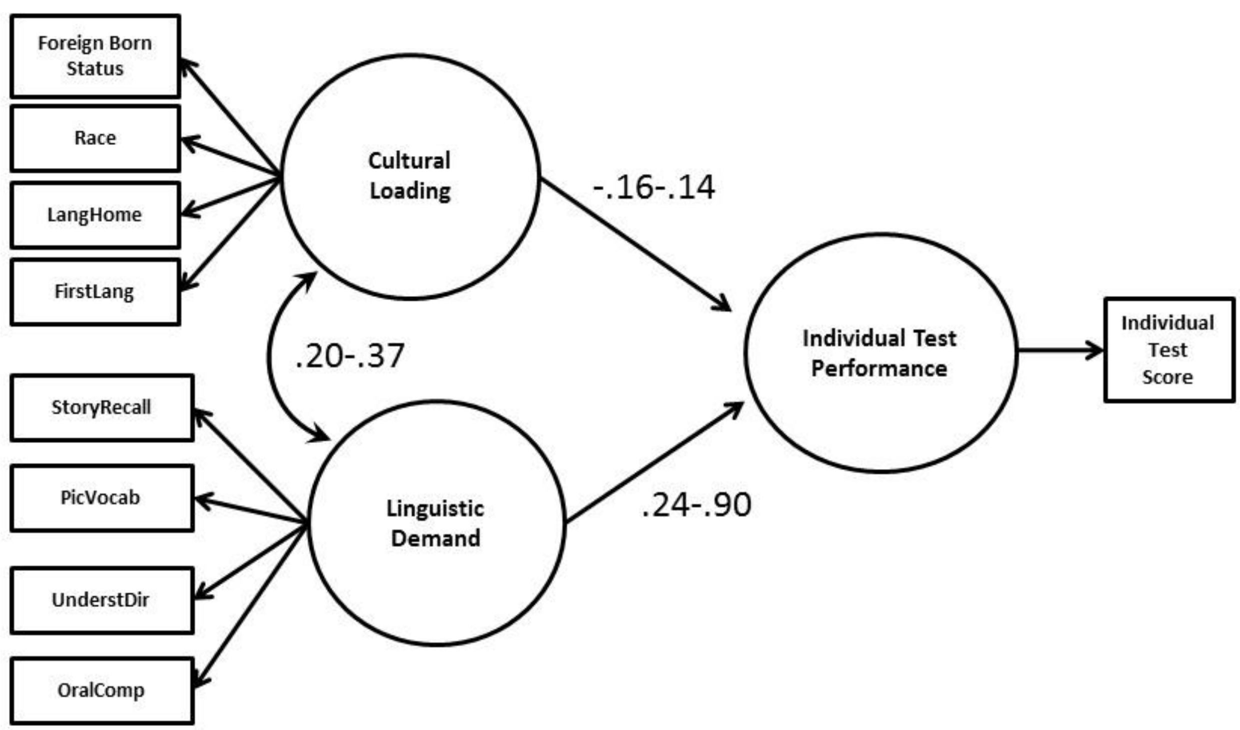

Figure 10. Path coefficient ranges from latent variable structural equation models for Phase 1, ages 15 to 18.

Variance Explained. Table 3 summarizes the variance in Individual Test Performance accounted for by the two latent variables (Cultural Loading and Linguistic Demand). In general, across age groups, there was a great deal of disparity in the variance explained by the variables. This is consistent with the strength of the association observed in the values from the previously discussed path coefficients.

Table 3

Phase 1 Variance Explained by Exogenous Variables for Test Variable by Model and Age Group

	Variance Explained		
Test from WJ-III Cognitive Battery	7 to 10	11 to 14	15 to 18
Analysis Synthesis	0.29	0.44	0.47
Auditory Attention	0.10	0.20	0.15
Auditory Working Memory	0.22	0.44	0.32
Concept Formation	0.67	0.71	0.67
Decision Speed	0.12	0.15	0.19
Delayed Recall Visual-Auditory Learning	0.39	0.32	0.37
General Information	0.71	0.85	0.86
Incomplete Words	0.13	0.31	0.23
Memory for Words	0.18	0.32	0.23
Numbers Reversed	0.17	0.26	0.3
Pair Cancellation	0.17	0.11	0.11
Picture Recall	0.02	0.06	0.10
Planning	0.07	0.12	0.11
Retrieval Fluency	0.22	0.22	0.28
Rapid Picture Naming	0.16	0.07	0.16
Sound Blending	0.25	0.32	0.35
Spatial Relations	0.08	0.16	0.16
Visual-Auditory Learning	0.40	0.37	0.41

Visual Matching	0.13	0.15	0.16
Verbal Comprehension	0.79	0.86	0.81

Summary

In general there was adequate model fit to make an interpretation of the path coefficients across the individual tests and three age groups. The results of the analysis suggest that only a few individual tests demonstrated a relationship strong enough to be considered *high* in Linguistic Demand. Overall, across all tests and age groups, Cultural Loading—as represented in this study—did not have significant paths. Despite not demonstrating a relationship to the latent variable Cultural Loading, the majority of the individual tests demonstrated at least a moderate relationship with the latent variable Linguistic Demand. Thus, Phase 2 aimed to provide further information on the relationship between Linguistic Demand and individual test performance by further specifying the contribution of receptive and expressive language when attempting to model the influence of overall Linguistic Demand.

Phase 2 – Expressive Language

Model Fit. The researcher evaluated model fit to determine how consistent the observed data were with the proposed model. Dividing the Chi-Square value by the degrees of freedom for the model provides a rough estimate of how well the data fit the model. Based on this estimate, the data appeared to fit relatively well across tests and age groups. In general, the values of this calculation across tests and age groups ranged from around 4 to 6. However, due to the influence of sample size on calculating Chi-Square statistics, the researcher again gathered additional fit information using RMSEA, CFI, and NFI values. Overall, model fit was relatively good across models for individual tests and for all three age groups. However, one individual test – Verbal Comprehension – demonstrated apparent problems across age groups. The overall fit based on the χ^2 divided by 2 method yielded values around 11, which were 3 to 4 times the values for all the other individual tests. However, the other fit indices (RMSEA, CFI, and NFI) were satisfactory. For information on specific fit indices, see Appendix C.

Table 4

Revised Fit Indices for Models with Problematic Factors Loadings in the Measurement Model

		Nested Model Comparison				Comparative Fit
Age Group	Test	$\chi2$	df	$\chi2$/df	NFI	CFI
7 to 10	Picture Recall	39.381	13	3.029	0.961	0.973
	Planning	37.195	13	2.861	0.963	0.975
	Spatial Relations	42.459	13	3.266	0.958	0.970
	Verbal Comprehension	133.536	13	10.272	0.923	0.929
11 to 14	Auditory Attention	41.986	13	3.230	0.969	0.978
	Decision Speed	50.342	13	3.872	0.961	0.971
	Numbers Reversed	57.679	13	4.437	0.957	0.966
	Pair Cancellation	44.47	13	3.421	0.966	0.975
	Picture Recall	48.118	13	3.701	0.963	0.972
	Planning	51.715	13	3.978	0.960	0.970
	Rapid Picture Naming	51.53	13	3.964	0.960	0.970
	Spatial Relations	51.135	13	3.933	0.961	0.971
	Visual Matching	53.722	13	4.132	0.959	0.968
15 to 18	Pair Cancellation	71.746	13	5.519	0.912	0.925
	Picture Recall	75.556	13	5.812	0.908	0.921
	Planning	63.564	13	4.890	0.922	0.935
	Spatial Relations	68.221	13	5.248	0.92	0.933
	Visual Matching	76.037	13	5.849	0.908	0.921

Verbal Comprehension	144.342	13	11.103	0.894	0.901

Path Coefficients. Results were relatively consistent across age groups. However, the output evinced a number of problematic path coefficients in the measurement model for a varying number of individual test models across age groups. The problematic path in the structural model was from Expressive Language to Individual Test Performance. A high level of colinearity between Story Recall and Picture Vocabulary was the likely cause. As a result, the researcher made a post-hoc adjustment to the model by fixing the path from Picture Vocabulary to Expressive Language to a value of 1 by setting the variance for its residual to a value of 0. As a result, this part of the measurement model would only estimate the parameter for Story Recall. All of the flagged models are bolded in the Tables seen in Appendix D. In addition, Table 5 summarizes the revised path coefficients for models with the fixed parameter.

Table 5

Revised Path Coefficients for Models with Problematic Factors Loadings in the Measurement Model

		Path		R^2	
Age Group	Test	Cultural Loading → Language	Language → Test	Individual Test Performance	Language
7 to 10	Picture Recall	0.30	0.02	0.00	0.09
	Planning	0.30	0.16	0.03	0.09
	Spatial Relations	0.30	0.15	0.02	0.09
	Verbal Comprehension	0.30	0.81	0.65	0.09
11 to 14	Auditory Attention	0.37	0.34	0.11	0.14
	Decision Speed	0.37	0.18	0.03	0.14
	Numbers Reversed	0.37	0.28	0.08	0.14

	Pair Cancellation	0.37	0.16	0.03	0.14
	Picture Recall	0.37	0.14	0.02	0.14
	Planning	0.37	0.20	0.04	0.14
	Rapid Picture Naming	0.37	0.15	0.02	0.14
	Spatial Relations	0.37	0.28	0.08	0.14
	Visual Matching	0.37	0.14	0.02	0.14
15 to 18	Pair Cancellation	0.36	0.08	0.01	0.13
	Picture Recall	0.36	0.16	0.03	0.13
	Planning	0.36	0.15	0.02	0.13
	Spatial Relations	0.36	0.29	0.08	0.13
	Visual Matching	0.36	0.16	0.02	0.13
	Verbal Comprehension	0.36	0.85	0.73	0.13

In this set of models run during Phase 2, the path coefficients from Cultural Loading to Expressive Language were relatively consistent within a group, but varied slightly from one age group to another. The path coefficients for the age group of 7 to 10 ranged from .25 to .39. As seen in Table D1.a (Appendix D), the individual tests that had relatively high values for the path coming from the latent variable Expressive Language and going to Individual Test Performance were General Information ($r = .75$) and Verbal Comprehension ($r = .81$). There were nine models that had paths from Expressive Language to Individual Test Performance that were not significant (less than $r = .30$). The path coefficients for ages 11 to 14 going from Cultural Loading to Expressive Language ranged from .24 to .39. The path coefficients from Expressive Language to Individual Test Performance ranged from .14 to .99 for the 11 to 14 age group. As seen in Table D2.a (Appendix D), the individual tests that had relatively high values for the path coming from the latent variable Expressive Language and going to Individual Test Performance, again, were General Information ($r = .94$) and Verbal Comprehension ($r =$

.99). The individual tests that had path coefficients that did not meet the threshold for significance (i.e. path coefficient less than .30) were Picture Recall (*r = .14*) and Planning (*r = .20*).

The path coefficients for ages 11 to 14 going from Cultural Loading to Expressive Language ranged from .28 to .41 for ages 15 to 18. The coefficients from Expressive Language to Individual Test Performance ranged from .08 to .85 for the 15 to 18 age group. As seen in Table D3.a (Appendix D), the individual tests that had relatively high values for the path coming from the latent variable Expressive Language and going to Individual Test Performance were Verbal Comprehension (*r = .85*) and General Information (*r = .94)*. There were six models that had paths from Expressive Language to Individual Test Performance that were not significant (less than *r* = .30).

Variance Explained. Table 6 summarizes the variance in Expressive Language accounted for by Cultural Loading, as well as the variance in Individual Test Performance accounted for by the latent variable Expressive Language. In general, across age groups, there was a great deal of disparity in the variance explained by the variables. This is consistent with the strength of the association observed in the values from the previously discussed path coefficients.

Table 6

Phase 2 Expressive Language Variance Explained by Exogenous Variables for Test Variable by Model and Age Group

	Variance Explained		
Test from WJ-III Cognitive Battery	7 to 10	11 to 14	15 to 18
Analysis Synthesis	0.12	0.17	0.26

Auditory Attention	0.05	0.11	0.10
Auditory Working Memory	0.16	0.25	0.23
Concept Formation	0.25	0.31	0.36
Decision Speed	0.02	0.03	0.04
Delayed Recall Visual-Auditory Learning	0.34	0.16	0.27
General Information	0.75	0.88	0.89
Incomplete Words	0.09	0.26	0.20
Memory for Words	0.12	0.25	0.14
Numbers Reversed	0.03	0.08	0.16
Pair Cancellation	0.04	0.03	0.01
Picture Recall	0.00	0.02	0.03
Planning	0.03	0.04	0.02
Retrieval Fluency	0.26	0.13	0.28
Rapid Picture Naming	0.15	0.02	0.10
Sound Blending	0.21	0.24	0.30
Spatial Relations	0.02	0.08	0.08
Visual-Auditory Learning	0.34	0.29	0.36
Visual Matching	0.03	0.02	0.02
Verbal Comprehension	0.65	0.98	0.73

Phase 2 – Receptive Language

Model Fit. The Researcher again tested the to determine if there was an adequate fit to the data. Dividing the Chi-Square value by the degrees of freedom for the model provides a rough estimate of how well the data fit the model. Based on this estimate, the data appeared to fit relatively well across tests and age groups. In general, the values of this calculation across tests and age groups ranged from approximately 2 to 7. However, the results were less consistent across age groups than those observed for Expressive Language—the minimum and maximum values for the ranges within age group tended to decrease as the age range of the groups increased, which suggests that the model fit was better for older students.

The researcher again gathered additional fit information using RMSEA, CFI, and NFI values. Overall, model fit was relatively good across models for individual tests and for all three age groups. As seen in Appendix C, the fit indices (RMSEA, CFI, and NFI) were satisfactory, with none of the models from the test or cross-validation analyses for all age groups and tests having NFI or CFI values below the threshold value of .90 or RMSEA values above .10.

Table 7

Revised Fit Indices for Models with Problematic Factors Loadings in the Measurement Model

		Nested Model Comparison				Comparative Fit
Age Group	Test	$\chi2$	Df	$\chi2$/df	NFI	CFI
7 to 10	Decision Speed	136.043	13	10.465	0.876	0.885
	Pair Cancellation	194.008	13	14.924	0.832	0.839
11 to 14	Decision Speed	133.542	13	10.272	0.905	0.912
15 to 18	Decision Speed	145.711	13	11.209	0.847	0.857
	Picture Recall	54.877	13	4.221	0.936	0.950
	Visual Matching	128.983	13	9.922	0.862	0.872

Path Coefficients. Results were relatively consistent across age groups. However, as was the case for Expressive Language, the analyses produced a number of problematic path coefficients in the measurement model for a varying number of individual test models across age groups. The problematic path in the structural model was from Receptive Language to Individual Test Performance. A high level of colinearity between Understanding Directions and Oral Comprehension was the likely cause. As a result, the researcher made post-hoc adjustments to the model by fixing the path for Oral

Comprehension to a value of 1 and by setting the variance for its residual to a value of 0. Consequently, this part of the measurement model would only estimate the parameter for Understanding Directions. All of the flagged models are bolded in the Tables seen in Appendix D. In addition, Table 8 summarizes the revised path coefficients for models with the fixed parameter.

Table 8

Revised Path Coefficients for Models with Problematic Factors Loadings in the Measurement Model

		Path		R^2	
Age Group	Test	Cultural Loading → Language	Language →Test	Individual Test Performance	Language
7 to 10	Decision Speed	0.20	0.16	0.03	0.04
	Pair Cancellation	0.20	0.15	0.02	0.04
11 to 14	Decision Speed	0.25	0.17	0.03	0.06
15 to 18	Decision Speed	0.18	0.21	0.04	0.03
	Picture Recall	0.19	0.16	0.03	0.03
	Visual Matching	0.19	0.22	0.05	0.03
	Decision Speed	0.18	0.20	0.04	0.03

In this set of models run during this part of Phase 2, the path coefficients from Cultural Loading to Receptive Language were relatively consistent within a group, but varied slightly from one age group to another. The path coefficients for the age group of 7 to 10 ranged from .03 to .34. As seen in Table D1.b (Appendix D), the individual tests that had relatively high values for the path coming from the latent variable Receptive Language and going to Individual Test Performance were Concept Formation (r = .85), General Information ($r = .86$) and Verbal Comprehension ($r = .96$). The individual tests that had path coefficients that did not meet the threshold for significance (i.e. path coefficient less than .30) were Picture Recall (*r* = .10) and Rapid Picture Naming (*r* = *.26*).

The path coefficients for ages 11 to 14 going from Cultural Loading to Receptive Language ranged from .16 to .36. The path coefficients from Receptive Language to Individual Test Performance ranged from .23 to .94 for the 11 to 14 age group. As seen in Table D2.b (Appendix D), the individual tests that had relatively high values for the path coming from the latent variable Receptive Language and going to Individual Test Performance, again, were Concept Formation (*r* = .83), General Information (*r = .91)* and Verbal Comprehension (*r = .94*). The individual tests that had path coefficients that did not meet the threshold for significance (i.e. path coefficient less than .30) were Picture Recall (*r = .23*) and Rapid Picture Naming (*r = .25*).

The path coefficients for ages going from Cultural Loading to Expressive Language ranged from .10 to .31 for ages 15 to 18. The coefficients from Expressive Language to Individual Test Performance ranged from .32 to .94 for the 15 to 18 age group. As seen in Table D3.b (Appendix D), the individual tests that had relatively high values for the path coming from the latent variable Expressive Language and going to Individual Test Performance were Concept Formation (*r* = .81), General Information (*r = .89)* and Verbal Comprehension (*r = .94*). No models had paths from Expressive Language to Individual Test Performance that were not significant (less than *r* = .30) for this age group.

Variance Explained. Table 9 summarizes the variance in Receptive Language accounted for by Cultural Loading, as well as the variance in Individual Test Performance accounted for by the latent variable Receptive Language. In general, across age groups, there was a great deal of disparity in the variance explained by the variables. This is consistent with the strength of the association observed in the values from the previously discussed path coefficients.

Table 9

Phase 2 Receptive Language Variance Explained by Exogenous Variables for Test Variable by Model and Age Group

	Variance Explained		
Test from WJ-III Cognitive Battery	7 to 10	11 to 14	15 to 18
Analysis Synthesis	0.39	0.46	0.51
Auditory Attention	0.16	0.24	0.17
Auditory Working Memory	0.30	0.42	0.36
Concept Formation	0.73	0.69	0.66
Decision Speed	0.03	0.03	0.04
Delayed Recall V-AL	0.45	0.33	0.36
General Information	0.73	0.82	0.8
Incomplete Words	0.10	0.27	0.19
Memory for Words	0.19	0.33	0.24
Numbers Reversed	0.23	0.30	0.37
Pair Cancellation	0.02	0.15	0.03
Picture Recall	0.01	0.05	0.05
Planning	0.11	0.11	0.17
Retrieval Fluency	0.23	0.26	0.28
Rapid Picture Naming	0.07	0.06	0.16
Sound Blending	0.27	0.32	0.33
Spatial Relations	0.12	0.18	0.16
Visual-Auditory Learning	0.41	0.37	0.41
Visual Matching	0.11	0.22	0.04
Verbal Comprehension	0.93	0.88	0.89

Note. V-AL = Visual-Auditory Learning

Summary

In general, there was adequate model fit to make an interpretation of the path coefficients across the individual tests and age groups. There appears to be convergent evidence, regardless of whether one takes expressive or receptive language into account, that there are generally two tests that have a high Linguistic Demand. As seen in the previous analyses, only two individual tests demonstrate a relationship of a strength that would be considered *high* in Linguistic Demand based on their path coefficients and the variance explained by the variables identified in the models. In conclusion, the increased specificity of the models tested in Phase 2 do provide some nuances with regard to tests of moderate Linguistic Demand overall (as seen in Phase 1), which may be of use to practitioners. However, the two Phase 1 tests that imposed high linguistic demands did not tend to demonstrate a significant difference in terms of their relationship with either expressive or receptive language. This suggests that they are equally loaded on these two parameters and that the general loading from Phase 1 should be sufficient in informing decision-making in test selection and future test development.

Chapter 4 - Discussion

Flanagan and Ortiz (2001) originally developed the C-LIM to help with test selection and interpretation. Practitioners are in need of this proactive approach, which helps to minimize the potential confounds related to the influences of Cultural Loading and Linguistic Demand by using the C-LIM as a guide in test selection *and* interpretation. This is particularly valuable when one considers the great deal of diversity that exists in schools across the United States. In its original development, the C-LIM was based mostly on expert opinion. When examining the classification more closely, some patterns emerged. For example, findings indicated that the most linguistically demanding tests were all tests of Comprehension-Knowledge (*Gc*) and Fluid Reasoning (*Gf*). These tests are: Concept Formation (R^2 = .71; *Gf*), General Information (R^2 = .85; *Gc*) and Verbal Comprehension (R^2 = .86; *Gc*). These tests demonstrated high Linguistic Demand in the C-LIM as well. Tests with particularly low Linguistic Demand tended to measure domains of Visual-spatial processing (*Gv*) and Processing speed (*Gs*): Picture Recognition (R^2 =.06; Gv), Planning (R^2 = .12; *Gv*), Rapid Picture Naming (R^2 = .07; *Gs*) and Spatial Relations (R^2 = .16; *Gv*). This appears to be in line with the general trends that would be expected, given the pattern of predicted performance assumed in the original C-LIM classifications (Flanagan & Ortiz, 2001). The original classification appears to require some modification, however, given the empirical evidence provided in this study.

Cultural Loading

Across both phases of the study, Cultural Loading, as captured in the latent variable in the model, did not have significant relationships to the performance on any of the individual tests included in the standard or extended batteries of the WJ-III COG battery. This suggests that a student's cultural background, as modeled here, may not have a significant impact on the student's performance on the tests, as was previously suggested in the C-LIM. It could be argued that the latent variable of Cultural Loading, as modeled in this study, does not capture a student's cultural background with enough specificity to see a significant effect. In other words, the relationship between Cultural Loading and individual test scores may exist, but this particular study's analyses did not demonstrate the relationship.

Conversely, the modeling of Cultural Loading as seen in this study may reflect what practitioners often interpret as observable factors that contribute to their decision-making in test selection and interpretation in practice. For example, it is likely that practitioners will consider, race, first language, language at home, and immigrant status when making decisions about test selection and interpretation. In fact, it is difficult to determine whether they consider more information than these four factors when selecting tests or when making interpretations of test scores. Thus, if this is an accurate reflection of actual practice, then this study suggests that making judgments based on these variables about test selection and interpretation may be leading practitioners to be overly sensitive to these variables thought to influence tests scores. Further, the empirical evidence provided by this study helps inform practice by demonstrating that these factors are unreliable in making judgments about the extent to which they may affect the individual test performance of a given student, regardless of the aforementioned factors (e.g., race). However, the inferences drawn from these results should take into consideration a number of limitations.

Linguistic Demand

Overall, there appear to be only three to four tests that have a high degree of Linguistic Demand from ages seven to 18, which represent a large proportion of the school-aged population to which school psychologists administer the WJ-III COG. Phase 1 of this study yielded results that demonstrated partial support for the previous classifications presented on the axis of Linguistic Demand in the C-LIM. Further, the wide range in the path values for the relationships between Linguistic Demand and Individual Test Performance supports the categorization of levels of influence for individual tests, as suggested in the C-LIM. However, this study provided a re-classification of the tests from the WJ-III COG battery that is supported by empirical evidence and quantitative, albeit somewhat arbitrary, cut-points for the three levels of influence (i.e. low, moderate, high). This revised classification could serve as an evidence-based guideline for practitioners in the selection and interpretation of the results from administering the WJ-III COG.

Phase 2 of this study provided information on the relative contributions of expressive language ability and receptive language ability in influencing performance on the 20 tests included in the WJ-III COG. The results of this additional step suggest that there are noteworthy differences in the contribution of each of these aspects of language ability. In general, across age groups, the receptive language ability of a student being administered any of the 20 tests from the WJ-III COG battery will provide an influence on test performance that is equal to (within .05) or greater than his or her expressive language ability. This piece of empirical evidence supports recommendations in assessing a student's academic language ability before proceeding with test administration and interpretation because his or her expressive language may not be representative of their academic language abilities (Schon, Shaftel, & Markham, 2008).

Despite the differences in the relative influences of receptive and expressive language on individual test performance, the proposed re-classification of the WJ-III COG tests based on the results of Phase 1 appear to be maintained, since there is very little to no difference between the path coefficients for Phase 1 and the path coefficients for receptive language in Phase 2 (see Table 10). This would suggest that the general re-classification based on the Phase 1 results can be used as a reference for students who have a lower receptive language or those that have a limited English language proficiency, which can be categorized in a number of ways and using a number of measures. The only exceptions to this general observation were the tests of Decision Speed, Pair Cancellation, and Visual Matching, which have little to no language component. Furthermore, the combined and receptive language path coefficients are generally in the low range, which would lead to an accurate interpretation of their C-LIM re-classifications, regardless of whether their combined or receptive path coefficient was used in the selection or interpretation of these tests.

Table 10

Comparison of Receptive and Combined (Receptive and Expressive) Path Coefficients from Linguistic Demand to Individual Test Performance

	Age Groups					
	7 to 10		11 to 14		15 to 18	
Test	C	R	C	R	C	R
Analysis Synthesis	0.55	0.63	0.69	0.68	0.71	0.71
Auditory Attention	0.31	0.40	0.45	0.49	0.24	0.41
Auditory Working Memory	0.48	0.55	0.70	0.65	0.55	0.60
Concept Formation	0.82	0.85	0.86	0.83	0.83	0.81
Decision Speed	0.35	**0.16**	0.41	**0.17**	0.45	**0.21**
DR Visual-Auditory Learning	0.64	0.67	0.59	0.57	0.63	0.60
General Information	0.79	0.86	0.91	0.91	0.90	0.89
Incomplete Words	0.25	0.32	0.40	0.52	0.42	0.44
Memory for Words	0.44	0.44	0.55	0.58	0.47	0.49
Numbers Reversed	0.41	0.48	0.53	0.55	0.55	0.61
Pair Cancellation	0.42	**0.15**	0.35	0.38	0.32	**0.16**
Picture Recall	0.14	0.10	0.26	0.23	0.33	**0.22**
Planning	0.26	0.33	0.36	0.34	0.34	0.41
Retrieval Fluency	0.46	0.48	0.47	0.51	0.52	0.53
Rapid Picture Naming	0.32	0.26	0.25	0.25	0.38	0.40
Sound Blending	0.47	0.52	0.57	0.57	0.58	0.57
Spatial Relations	0.28	0.35	0.42	0.42	0.41	0.40
Visual-Auditory Learning	0.65	0.64	0.62	0.61	0.66	0.64
Visual Matching	0.37	0.46	0.41	0.46	0.41	**0.20**
Verbal Comprehension	0.90	0.96	0.94	0.94	0.90	0.94

Note. C = Combined (Receptive and Expressive), R = Receptive. Values in bold indicates that cross-validation results were replaced with values from the revised structural models with fixed measurement model parameters.

In summary, the results of this study suggest that the degree of Linguistic Demand varies a great deal between the tests included in the WJ-III COG. Moreover, the receptive language abilities of a student appear to be an important consideration in test selection and interpretation. Finally, the results of this study provide empirical evidence to support a re-classification of the tests presented in the current version of the C-LIM (Rhodes, Ochoa, & Ortiz, 2005).

Age Group Differences

The results of this study suggest a significant modification to the original C-LIM classification of the tests from the WJ-III COG battery. This trend was consistent across all three age groups (i.e., 7-10, 11-14, and 15-18), with fewer tests classified as high and a greater number of tests classified as moderate in Linguistic Demand than was proposed by Rhodes, Ochoa, and Ortiz (2005). Further, there were a greater proportion of tests classified as moderate in Linguistic Demand as student age increased. Conversely, there were fewer tests classified as low in Linguistic Demand with an increase in age, with only a single test classified as low in Linguistic Demand for ages 15 to 18.

This trend may reflect an increase in the difficulty of the vocabulary required to provide correct answers to items on certain tests that are typically only given to older students. In other words, as students get older, they are likely to be exposed to a greater number of items with increasing difficulty, given the structure of the WJ-III COG and its use of ceiling rules to guide test administration. These results can be compared to those found in Cormier, McGrew, and Evans (2011), where only the Linguistic Demand of test directions was considered. Taken together, the results of these two studies suggest that certain tests, particularly those classified as high in Linguistic Demand, should be used with caution when testing CLD students with characteristics that may attenuate their performance. For example, those categorized as high only in Linguistic Demand in Cormier et al. (2011) could be avoided when using a translator to administer the WJ-III COG, because the translation of a test with a high Linguistic Demand in the test directions is more likely to be misinterpreted or break standardization if the directions are overly complex or verbose. Further, the results of Cormier et al. and the Phase 2 of this study could be particularly important when testing CLD students who are English Language Learners, due to the variation in their level of receptive and expressive language ability.

7 to 10		11 to 14		15 to 18		
Test	Path Coefficient	Test	Path Coefficient	Test	Path Coefficient	C-LIM
Picture Recognition	0.14	Rapid Picture Naming	0.25	Auditory Attention	0.24	Spatial Relations
Incomplete Words	0.25	Picture Recognition	0.26	Pair Cancellation	0.32	Picture Recognition
Planning	0.26	Pair Cancellation	0.35	Picture Recognition	0.33	Planning
Spatial Relations	0.28	Planning	0.36	Planning	0.34	Pair Cancellation
Auditory Attention	0.31	Incomplete Words	0.4	Rapid Picture Naming	0.38	Visual Matching
Rapid Picture Naming	0.32	Decision Speed	0.41	Visual Matching	0.41	Numbers Reversed
Decision Speed	0.35	Visual Matching	0.41	Spatial Relations	0.41	Visual Auditory Learning
Visual Matching	0.37	Spatial Relations	0.42	Incomplete Words	0.42	DR-Visual Auditory Learning
Numbers Reversed	0.41	Auditory Attention	0.45	Decision Speed	0.45	Retrieval Fluency
Pair Cancellation	0.42	Retrieval Fluency	0.47	Memory for Words	0.47	Rapid Picture Naming
Memory for Words	0.44	Numbers Reversed	0.53	Retrieval Fluency	0.52	Concept Formation
Retrieval Fluency	0.46	Memory for Words	0.55	Numbers Reversed	0.55	Analysis Synthesis
Sound Blending	0.47	Sound Blending	0.57	Auditory Working Memory	0.55	Auditory Working Memory
Auditory Working Memory	0.48	DR-Visual Auditory Learning	0.59	Sound Blending	0.58	Memory for Words
Analysis Synthesis	0.55	Visual Auditory Learning	0.62	DR-Visual Auditory Learning	0.63	Incomplete Words
DR-Visual Auditory Learning	0.64	Analysis Synthesis	0.69	Visual Auditory Learning	0.66	Sound Blending
Visual Auditory Learning	0.65	Auditory Working Memory	0.7	Analysis Synthesis	0.71	Auditory Attention
General Information	0.79	Concept Formation	0.86	Concept Formation	0.83	Decision Speed
Concept Formation	0.82	General Information	0.91	General Information	0.9	Verbal Comprehension
Verbal Comprehension	0.9	Verbal Comprehension	0.94	Verbal Comprehension	0.9	General Information

Figure 11. Comparison of current and revised C-LIM test classifications for tests from the Woodcock-Johnson Test of Cognitive Abilities, Third Edition.

Limitations

There are a few noteworthy limitations that one should take into consideration when drawing inferences from the results of this study. First, the sample used was the Woodcock Johnson Third Edition Normative Update norming data. These data present a certain limitation: There is an under-representation of immigrant populations and English Language Learner students in the norming sample. Therefore, this study was unable to determine the relationships among the variables for these specific populations. However, this study did account for diversity to some degree and used a broad range of linguistic ability to model the construct of Linguistic Demand. Future research could address this limitation by attempting to use similar models with a sample that has a larger proportion of CLD students.

This study also had an additional limitation: the difficulty in modeling the construct of Cultural Loading, which influenced, to a relatively negligible degree, the endogenous variable of Individual Test Performance. This limitation was due to the nature of the dataset used, which only had a limited number of variables that could be selected to model this latent variable. However, it is unclear how the variable could be improved in future research because the variable of culture would be difficult to capture due to the level of diversity that occurs within racial or ethnic categories.

Implications for Research

This research provided empirical evidence to suggest revisions should be made to the C-LIM as it was first proposed and is currently in use by practitioners. Due to the significant number of changes that were made, future research should aim to gather empirical evidence to either support or modify the original classifications that have been made for other major batteries of cognitive abilities, such as the *Kaufman Assessment Battery for Children* and the *Wechsler Intelligence Scale Children*. This research also provides information for the development of future measures of cognitive abilities. For example, the simplification of test directions to reduce the impact of a student's potentially limited receptive language abilities would be beneficial to the validity of future tests. In addition, further studies could be conducted to gather convergent empirical evidence to support the re-classifications presented herein. Moreover, the re-classification proposed by this research should be tested in ways similar to those seen in Kranzler et al. (2011) to further validate the re-classification.

Implications for Practice

School Psychology. Due to the limited influence of culture, as measured in this study, implications for practice are limited to the C-LIM dimension of Linguistic Demand. Despite age differences in cognitive abilities, and consequently the way in which students interact with test items, most tests appeared to have similar categorizations across age groups. This consistency makes the use of this re-classification easier for practitioners, as age-specific matrices may not be needed to guide general decision-making about test selection and interpretation. Moreover, even the tests that went from one categorical label to another had a tendency to be very close in their r-squared values and relative rank order in the classification. For example, Analysis

Synthesis and Auditory Working Memory tended to stay in the higher end of the moderate range or the lower end of the high range across age groups (see Figure 11). Further, based on this empirical evidence, practitioners should use and interpret those tests classified as *high* linguistic demand with caution, especially when testing CLD students. Conversely, practitioners should also proceed with caution with those listed in the *moderate* range, although their influence on attenuating test scores should be less than the effect of those classified as *high*. It should be noted that this research does not provide a numeric value as suggested in Flanagan and Ortiz (2001). This may be another area for which future research could provide empirical support.

Ethical Standards and Educational Policy. The results of this study demonstrate the need for mandates that require practices to be supported by empirical evidence. This is particularly important considering the history of legal and ethical dilemmas involving inappropriate assessment practices being used with students from culturally and linguistically diverse backgrounds. Currently, ethical standards appear to be more stringent than special education laws outlining assessment practices that are permissible when evaluating CLD students for special education services because they provide a more dynamic approach to the nature of testing. For example, the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) standard 12.16 states that "test interpretations should not imply that empirical evidence exists for a relationship among specific test results, prescribed interventions, and desired outcomes, unless empirical evidence is available for populations similar to those representative of the examinee (p.134)." Thus, standard 12.16 suggests that the C-LIM should not be used in test interpretation, as no empirical evidence supports its use in its current form. As such, special education law should extend general guidelines on testing to require not only reliability and validity evidence for the assessments used by practitioners, but also specify the requirement of having empirical evidence to support the test interpretations that are reported by practitioners.

School Administrators. The potential impact of increasing the specificity of the requirements for empirical evidence to support test interpretation practices is that school administrators as well as district personnel are more likely to be aware of these requirements and engage in conversations with practitioners to ensure these practices are in places in their schools or districts. Ideally, the message that would be conveyed through laws specifying the need for empirical support of assessment interpretations is the need for comprehensive assessment practices—a message that could potentially be consistently conveyed not only from professional organizations, but from school and district administrators that may currently be unaware of this necessary component in making appropriate educational decisions for students.

Conclusion

The broader implications of this research are to promote comprehensive evaluation processes in schools, particularly when working with CLD students. The implications of this study go beyond a re-classification of the C-LIM to emphasizing the importance of considering a student's linguistic background and abilities prior to selecting, administering, and interpreting tests of cognitive abilities. Due to the relatively high influence of language for tests that load on General Intellectual Ability (e.g. Verbal Comprehension, Concept Formation, and General Information), it is necessary to consider a student's language ability throughout the evaluation process in order to ensure practitioners make appropriate decisions from assessment results (i.e. consequential validity). In addition, a student's expressive language ability (i.e. conversational proficiency) may not be an accurate representation of a student's receptive language abilities. Thus, it would be beneficial to gather information on a student's receptive language ability, due to its higher level of influence on individual test performance, relative to the contribution of expressive language.

Empirical evidence on the influences of culture and language when assessing the cognitive abilities of CLD students continues to accumulate. This gathered evidence will likely lead to an increase in the awareness of testing consideration for CLD students and improve practices of school psychologists. One of the areas that may continue to see improvements, particularly because of the significant results related to Linguistic Demand in this study, is the assessment of language proficiency in aiding school psychologists' decisions about test selection and interpretation. The results of this study may influence practice by encouraging school psychologists to move away from frequently using informal measures of linguistic ability through language samples and student interviews (Ochoa, Galarza, & Gonzalez, 1996) to more rigorous measures of language ability with an emphasis on receptive language. Regardless of the impact of this study, it appears that the degree of linguistic demand of tests will continue to be an important consideration in the assessment of cognitive abilities of CLD students.

References

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

Arlin, P.K. (1975). Cognitive development in adulthood: A fifth stage? *Developmental Psychology, 11*, 602-606.

Bracken, B. A., & Naglieri, J. A. (2003). Assessing diverse populations with nonverbal tests of general intelligence. *Handbook of Psychological and Educational Assessment of Children: Intelligence, Aptitude, and Achievement, 2*, 243-274.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since bias in mental testing. *School Psychology Quarterly*, *14*(3), 208-238.

Cormier, D.C., McGrew, K.S., Evans, J.J. (2011). Quantifying the "Degree of Linguistic Demand" in Spoken Intelligence Test Directions. *Journal of Psychoeducational Assessment*, *0*, 0-0. doi: 10.1177/0734282911405962

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

Flanagan, D. P., & Harrison, P. L. (2005). *Contemporary intellectual assessment: Theories, tests, and issues.* Guilford Press New York.

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

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). *Essentials of Cross-Battery Assessment with C/D ROM* (2nd ed.). New York: Wiley.

Fletcher, J. M., Francis, D. J., Rourke, B. P., Shaywitz, S. E., & Shaywitz, B. A. (1992). The validity of discrepancy-based definitions of reading disabilities. *Journal of Learning Disabilities, 25*(9), 555.

Fletcher, J.M., Francis, D.J., Shaywitz, S.E., Lyon, G.R., Foorman, B.R., Stuebing, K.K., & Shaywitz, B.A. (1998). Intelligent testing and the discrepancy model for children with learning disabilities. *Learning Disabilities Research & Practice, 13*(4), 186-203.

Francis, D. J., Espy, K. A., Rourke, B. P., & Fletcher, J. M. (1991). Validity of intelligence test scores in the definition of learning disability: A critical analysis. *Validity Issues in Learning Disabilities,* , 12–40.

Francis, D. J., Fletcher, J. M., Shaywitz, B. A., Shaywitz, S. E., & Rourke, B. P. (1996). Defining learning and language disabilities: Conceptual and psychometric issues with the use of IQ tests. *Language, Speech, and Hearing Services in Schools, 27*(2), 132.

Garcia, E., & Cuellar, D. (2006). Who are these linguistically and culturally diverse students? *The Teachers College Record, 108*(11), 2220-2246.

Hansen, A., Cormier, D.C., Lim, B., & McGrew, K.S. (March, 2010). *Cultural Loading and Linguistic Demand on the WJ III NU*. Poster presented at the 2010 National Association of School Psychologists, Chicago, Illinois.

Harry, B., & Klingner, J. (2006). *Why are so many minority students in special education?* London, England: Teacher College Press.

Hosp, J.L., & Reschly, D.J. (2004). Disproportionate representation of minority students in special education: Academic, demographic, and economic predictors. *Exceptional Children, 70*(2), 185-199.

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

Klingner, J. K., Artiles, A. J., Kozleski, E., Harry, B., Zion, S., Tate, W., Duran, G. Z., & Riley, D. (2005). Addressing the disproportionate representation of culturally and linguistically diverse students in special education through culturally responsive educational systems. *Education Policy Analysis Archives, 13*, 43.

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.

McGrew, K.S. (2005). CHC theory of cognitive abilities. In D. P. Flanagan & P.L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, tests and issues* (pp. 136-181). New York: Guilford.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1-10.

McGrew, K.S., Keith, T.Z., Flanagan, D.P., Vanderwood, M. (1997) Beyond "g": The impact of "Gf-Gc" specific cognitive abilities research on the future use and interpretation of Intelligence test batteries in the schools. *School Psychology Review*, 26, 189-210.

McGrew, K., & Schrank, F., & Woodcock, W.R. (2007). Technical manual, woodcock-johnson III normative update.

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

Mertens, D.M. (2003). Consequential validity. In Lewis-Beck, M.S., Bryman, A., & Liao, T.F. (Eds.). The SAGE Encyclopedia of Social Science Research Methods. Thousand Oaks, CA: SAGE Publications.

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). St. John's University, New York, New York.

National Research Council (1982). *Minority students in special and gifted education.* Washington, DC: National Academy Press.

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

O'Connor, C., & Fernandez, S.D. (2006). Race, class, and disproportionality: Reevaluating the relationship between poverty and special education placement. *Educational Researcher, 35*(6), 6-11.

Ochoa, S.H., Galarza, A., & Gonzalez, D. (1996). An investigation of school psychologists' assessment practices of language proficiency with bilingual and limited-English-proficient students. *Assessment for Effective Intervention 21*(4) 17-36. doi: 10.1177/073724779602100402

Ortiz, A. A. (1997). Learning disabilities occurring concomitantly with linguistic differences. *Journal of Learning Disabilities, 30*(3), 321.

Ortiz, S.O., & Dynda, A.M. (2005). Use of intelligence tests with culturally and linguistically diverse populations. In D. P. Flanagan & P.L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, tests and issues* (pp. 545-556). New York: Guilford.

Padilla, A. M., & Perez, W. (2003). Acculturation, social identity, and social cognition: A new perspective. *Hispanic Journal of Behavioral Sciences, 25*(1), 35.

Reynolds, C. R. (2000). Why is psychometric research on bias in mental testing so often ignored? *Psychology, Public Policy, and Law, 6*(1), 144-150.

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

Salvia, J., & Ysseldyke, J. (2004). *Assessment in special education and inclusive education*. Boston: Houghton Mifflin Company

Salvia, J., Ysseldyke, J. & Bolt, S. (2010). *Assessment in special and inclusive education*. Boston, MA: Cengage.

Sayegh, L., & Lasry, J.-C. (1993). Immigrants' adaptation to Canada: Assimilation, acculturation and orthogonal cultural identification. *Canadian Psychology, 34,* 98-109.

Schon, J., Shaftel, J., & Markham, P. (2008). Contemporary issues in the assessment of culturally and linguistically diverse learners. *Journal of Applied School Psychology*, *24*, 163-189.

Stuebing, K.K., Fletcher, J.M., LeDoux, J.M., Lyon, G.R., Shaywitz, S.E., Lyon, G.R., & Shaywitz, B.A. (2002). *American Education Research Journal, 39*(2), 469-518.

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, New York, NY.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*(1), 29-37.

Verderosa, F.A. (2007). *Examining the effects of language and culture on the differential ability scales with bilingual preschoolers* (Unpublished doctoral dissertation). St. John's University, New York, NY.

Wasserman, J. D., & Tulsky, D. S. (2005). A history of intelligence assessment. In D. P. Flanagan & P.L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, tests and issues* (pp. 3-22). New York: Guilford.

Wechsler, D. (1939). *The measurement of adult intelligence.* Baltimore, MD: Williams & Wilkins.

Wellman, H.M., & Gelman, S.A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43*, 337-375.

Ysseldyke, J., Burns, M., Dawson, P., Kelley, B., Morrison, D., Ortiz, S., Rosenfield, S., & Telzrow, C. (2006). School psychology: A blueprint for training and practice III. Bethesda, MD*: National Association of School Psychologists.*

Ysseldyke, J.E., Vanderwood, M.L., & Shriner, J. (1997). Changes over the past decade in special education referral to placement probability: An incredibly reliable practice. *Assessment for Effective Intervention, 23*(1), 193-201.

Appendices

Appendix A – Phase 1 Fit Indices

Table A.1

Phase 1 Fit Indices for Ages 7 to 10

	Test					Cross-Validation				
	χ2	df	χ2/df	NFI	CFI	χ2	df	χ2/df	NFI	CFI
Analysis Synthesis	155.672	24	6.486	0.912	0.924	133.099	24	5.546	0.920	0.933
Auditory Attention	107.825	24	4.493	0.931	0.945	116.165	24	4.840	0.925	0.939
Auditory Working Memory	121.048	24	5.044	0.929	0.939	108.528	24	4.522	0.932	0.945
Concept Formation	219.161	24	9.132	0.892	0.902	252.621	24	10.526	0.877	0.886
Decision Speed	178.613	24	7.442	0.892	0.904	157.868	24	6.578	0.901	0.914
Delayed Recall Visual-Auditory Learning	116.915	24	4.871	0.928	0.941	104.184	24	4.341	0.937	0.950
General Information	118.423	24	4.934	0.947	0.957	131.929	24	5.497	0.946	0.936
Incomplete Words	91.536	24	3.814	0.944	0.958	110.005	24	4.584	0.930	0.944
Memory for Words	98.701	24	4.113	0.939	0.952	108.049	24	4.502	0.933	0.946
Numbers Reversed	125.492	24	5.229	0.923	0.936	127.405	24	5.309	0.920	0.934
Pair Cancellation	207.915	24	8.663	0.877	0.888	192.538	24	8.022	0.884	0.895
Picture Recall	95.333	24	3.972	0.938	0.952	100.206	24	4.175	0.933	0.948
Planning	103.442	24	4.310	0.933	0.947	105.388	24	4.391	0.931	0.945
Retrieval Fluency	128.284	24	5.345	0.924	0.936	113.011	24	4.709	0.930	0.943
Rapid Picture Naming	101.37	24	4.224	0.935	0.949	99.548	24	4.148	0.936	0.950
Sound Blending	118.994	24	4.958	0.931	0.943	125.669	24	5.236	0.925	0.938
Spatial Relations	118.994	24	4.958	0.93	0.944	109.206	24	4.550	0.929	0.943
Visual-Auditory Learning	117.285	24	4.887	0.931	0.944	100.582	24	4.191	0.941	0.954
Visual Matching	176.655	24	7.361	0.896	0.907	164.367	24	6.849	0.911	0.899
Verbal Comprehension	141.209	24	5.884	0.937	0.947	119.753	24	4.990	0.950	0.590

Table A.2

Phase 1 Fit Indices for Ages 11 to 14

	Test					Cross-Validation				
	$\chi 2$	df	$\chi 2$/df	NFI	CFI	$\chi 2$	df	$\chi 2$/df	NFI	CFI
Analysis Synthesis	126.233	24	5.260	0.915	0.928	150.531	24	6.272	0.926	0.937
Auditory Attention	100.113	24	4.171	0.921	0.936	107.311	24	4.471	0.944	0.955
Auditory Working Memory	104.464	24	4.353	0.920	0.933	100.986	24	4.208	0.949	0.960
Concept Formation	178.044	24	7.419	0.896	0.906	209.764	24	8.740	0.912	0.921
Decision Speed	164.264	24	6.844	0.904	0.918	145.55	24	6.065	0.925	0.936
Delayed Recall Visual-Auditory Learning	107.296	24	4.471	0.931	0.945	107.797	24	4.492	0.945	0.956
General Information	116.423	24	4.851	0.945	0.956	101.714	24	4.238	0.958	0.968
Incomplete Words	94.845	24	3.952	0.932	0.947	93.805	24	3.909	0.952	0.964
Memory for Words	102.227	24	4.259	0.918	0.931	100.332	24	4.181	0.95	0.961
Numbers Reversed	120.548	24	5.023	0.918	0.932	107.176	24	4.466	0.945	0.956
Pair Cancellation	158.313	24	6.596	0.888	0.900	118.005	24	4.917	0.938	0.949
Picture Recall	104.282	24	4.345	0.927	0.942	99.241	24	4.135	0.946	0.958
Planning	107.847	24	4.494	0.918	0.933	96.893	24	4.037	0.948	0.960
Retrieval Fluency	101.545	24	4.231	0.923	0.937	106.019	24	4.417	0.945	0.956
Rapid Picture Naming	96.644	24	4.027	0.926	0.940	96.727	24	4.030	0.947	0.959
Sound Blending	102.135	24	4.256	0.946	0.932	102.908	24	4.288	0.949	0.960
Spatial Relations	106.047	24	4.419	0.921	0.936	106.598	24	4.442	0.944	0.955
Visual-Auditory Learning	110.126	24	4.589	0.936	0.949	112.711	24	4.696	0.944	0.955
Visual Matching	149.049	24	6.210	0.893	0.906	150.435	24	6.268	0.923	0.933
Verbal Comprehension	149.618	24	6.234	0.938	0.948	104.018	24	4.334	0.96	0.968

Table A.3

Phase 1 Fit Indices for Ages 15 to 18

	Test					Cross-Validation				
	$\chi 2$	df	$\chi 2$/df	NFI	CFI	$\chi 2$	df	$\chi 2$/df	NFI	CFI
Analysis Synthesis	155.672	24	6.486	0.924	0.936	147.338	24	6.139	0.901	0.915
Auditory Attention	107.825	24	4.493	0.933	0.948	112.471	24	4.686	0.917	0.932
Auditory Working Memory	121.048	24	5.044	0.932	0.946	120.240	24	5.010	0.913	0.928
Concept Formation	219.161	24	9.132	0.907	0.918	179.952	24	7.498	0.896	0.907
Decision Speed	178.613	24	7.442	0.897	0.910	156.954	24	6.540	0.902	0.888
Delayed Recall Visual-Auditory Learning	116.915	24	4.871	0.931	0.945	141.526	24	5.897	0.901	0.915
General Information	118.423	24	4.934	0.940	0.951	135.000	24	5.625	0.927	0.939
Incomplete Words	91.536	24	3.814	0.938	0.952	108.781	24	4.533	0.920	0.936
Memory for Words	98.701	24	4.113	0.935	0.949	121.949	24	5.081	0.913	0.928
Numbers Reversed	125.492	24	5.229	0.923	0.937	132.911	24	5.538	0.908	0.922
Pair Cancellation	207.915	24	8.663	0.900	0.913	160.646	24	6.694	0.883	0.897
Picture Recall	95.333	24	3.972	0.930	0.944	107.539	24	4.481	0.918	0.934
Planning	103.442	24	4.310	0.927	0.941	122.804	24	5.117	0.908	0.923
Retrieval Fluency	128.284	24	5.345	0.935	0.949	121.256	24	5.052	0.914	0.929
Rapid Picture Naming	101.37	24	4.224	0.935	0.950	110.389	24	4.600	0.917	0.933
Sound Blending	118.994	24	4.958	0.935	0.949	125.902	24	5.246	0.914	0.928
Spatial Relations	106.047	24	4.419	0.930	0.945	119.004	24	4.959	0.912	0.928
Visual-Auditory Learning	117.285	24	4.887	0.932	0.945	115.933	24	4.831	0.921	0.935
Visual Matching	176.655	24	7.361	0.905	0.918	150.352	24	6.265	0.892	0.906
Verbal Comprehension	141.209	24	5.884	0.927	0.937	121.743	24	5.073	0.939	0.949

Appendix B – Phase 1 Path Coefficients

Table B.1

Phase 1 Results for Ages 7 to 10

	Test			Cross-Validation		
	Cultural Loading → Test Performance	*Linguistic Demand → Test Performance*	*Cultural Loading ←→ Linguistic Demand*	*Cultural Loading → Test Performance*	*Linguistic Demand → Test Performance*	*Cultural Loading ←→ Linguistic Demand*
Analysis Synthesis	-0.15	0.67	0.30	-0.17	0.55	0.24
Auditory Attention	-0.03	0.28	0.30	-0.02	0.31	0.24
Auditory Working Memory	-0.08	0.54	0.30	-0.04	0.48	0.24
Concept Formation	-0.11	0.82	0.26	-0.07	0.82	0.14
Decision Speed	-0.08	0.38	0.30	-0.09	0.35	0.23
Delayed Recall Visual-Auditory Learning	-0.07	0.56	0.31	-0.06	0.64	0.24
General Information	0.11	0.86	0.31	0.16	0.79	0.26
Incomplete Words	0.13	0.42	0.30	0.21	0.25	0.24
Memory for Words	0.00	0.41	0.30	-0.09	0.44	0.24
Numbers Reversed	-0.12	0.47	0.30	-0.18	0.41	0.23
Pair Cancellation	-0.11	0.41	0.30	-0.12	0.42	0.23
Picture Recall	-0.07	0.27	0.31	-0.01	0.14	0.24
Planning	-0.05	0.23	0.31	0.02	0.26	0.24
Retrieval Fluency	-0.01	0.51	0.30	0.04	0.46	0.23
Rapid Picture Naming	-0.03	0.36	0.31	0.17	0.32	0.24
Sound Blending	-0.01	0.53	0.31	0.08	0.47	0.24
Spatial Relations	-0.06	0.37	0.30	-0.10	0.28	0.24
Visual-Auditory Learning	-0.02	0.58	0.30	-0.14	0.65	0.24
Visual Matching	-0.15	0.45	0.31	-0.10	0.37	0.24
Verbal Comprehension	-0.12	1.01	0.30	-0.03	0.90	0.24

Table B.2

Phase 1 Results for Ages 11 to 14

	Test			Cross-Validation		
	Cultural Loading → Test Performance	*Linguistic Demand →Test Performance*	*Cultural Loading ←→ Linguistic Demand*	*Cultural Loading→ Test Performance*	*Linguistic Demand → Test Performance*	*Cultural Loading ←→Linguistic Demand*
Analysis Synthesis	-0.05	0.73	0.30	-0.19	0.69	0.30
Auditory Attention	-0.02	0.40	0.32	0.00	0.45	0.33
Auditory Working Memory	0.03	0.53	0.32	-0.21	0.70	0.33
Concept Formation	0.00	0.82	0.23	-0.07	0.86	0.22
Decision Speed	-0.19	0.53	0.30	-0.13	0.41	0.32
Delayed Recall Visual-Auditory Learning	-0.06	0.62	0.32	-0.15	0.59	0.32
General Information	0.03	0.89	0.36	0.04	0.91	0.34
Incomplete Words	0.15	0.45	0.32	0.28	0.40	0.31
Memory for Words	-0.01	0.55	0.31	0.05	0.55	0.33
Numbers Reversed	-0.09	0.53	0.31	-0.13	0.53	0.32
Pair Cancellation	-0.22	0.52	0.30	-0.13	0.35	0.32
Picture Recall	0.02	0.35	0.33	-0.16	0.26	0.33
Planning	-0.12	0.33	0.32	-0.07	0.36	0.33
Retrieval Fluency	-0.08	0.56	0.33	-0.01	0.47	0.32
Rapid Picture Naming	0.01	0.39	0.33	0.05	0.25	0.33
Sound Blending	-0.01	0.55	0.33	-0.01	0.57	0.33
Spatial Relations	-0.08	0.46	0.32	-0.06	0.42	0.32
Visual-Auditory Learning	-0.08	0.63	0.32	-0.04	0.62	0.32
Visual Matching	-0.17	0.49	0.31	-0.11	0.41	0.32
Verbal Comprehension	-0.05	1.02	0.38	-0.05	0.94	0.32

Table B.3

Phase 1 Results for Ages 15 to 18

Test	Test			Cross-Validation		
	Cultural Loading → Test Performance	*Linguistic Demand → Test Performance*	*Cultural Loading ←→ Linguistic Demand*	*Cultural Loading → Test Performance*	*Linguistic Demand → Test Performance*	*Cultural Loading ←→ Linguistic Demand*
Analysis Synthesis	-0.15	0.67	0.30	-0.16	0.71	0.23
Auditory Attention	-0.03	0.28	0.30	0.06	0.24	0.37
Auditory Working Memory	-0.08	0.54	0.30	0.04	0.55	0.24
Concept Formation	-0.11	0.82	0.26	-0.12	0.83	0.20
Decision Speed	-0.08	0.38	0.30	-0.07	0.45	0.24
Delayed Recall Visual-Auditory Learning	-0.07	0.56	0.31	-0.12	0.63	0.24
General Information	0.11	0.86	0.31	0.07	0.90	0.28
Incomplete Words	0.13	0.42	0.30	0.14	0.42	0.25
Memory for Words	0.00	0.41	0.30	0.01	0.47	0.24
Numbers Reversed	-0.12	0.47	0.30	0.00	0.55	0.24
Pair Cancellation	-0.11	0.41	0.30	0.02	0.32	0.24
Picture Recall	-0.07	0.27	0.31	-0.06	0.33	0.25
Planning	-0.05	0.23	0.31	-0.07	0.34	0.25
Retrieval Fluency	-0.01	0.51	0.30	0.02	0.52	0.25
Rapid Picture Naming	-0.03	0.36	0.31	0.05	0.38	0.25
Sound Blending	-0.01	0.53	0.31	0.04	0.58	0.26
Spatial Relations	-0.08	0.46	0.32	-0.05	0.41	0.25
Visual-Auditory Learning	-0.02	0.58	0.30	-0.10	0.66	0.24
Visual Matching	-0.15	0.45	0.31	-0.07	0.41	0.24
Verbal Comprehension	-0.12	1.01	0.30	0.00	0.90	0.24

Appendix C – Phase 2 Fit Indices

Table C.1a

Phase 2 Expressive Language Fit Indices for Ages 7 to 10

	Test					Cross-Validation				
	$\chi 2$	df	$\chi 2$/df	NFI	CFI	$\chi 2$	df	$\chi 2$/df	NFI	CFI
Analysis Synthesis	81.365	12	6.780	0.927	0.936	58.792	12	4.899	0.945	0.955
Auditory Attention	63.201	12	5.267	0.938	0.948	55.038	12	4.587	0.947	0.957
Auditory Working Memory	67.994	12	5.666	0.936	0.946	53.52	12	4.460	0.950	0.960
Concept Formation	73.666	12	6.139	0.937	0.946	66.48	12	5.540	0.942	0.952
Decision Speed	58.761	12	4.897	0.942	0.952	46.599	12	3.883	0.954	0.965
Delayed Recall Visual-Auditory Learning	58.175	12	4.848	0.944	0.954	55.24	12	4.603	0.950	0.960
General Information	99.769	12	8.314	0.936	0.943	73.243	12	6.104	0.950	0.957
Incomplete Words	63.064	12	5.255	0.942	0.952	58.955	12	4.913	0.945	0.955
Memory for Words	58.764	12	4.897	0.944	0.955	56.576	12	4.715	0.948	0.958
Numbers Reversed	62.465	12	5.205	0.940	0.950	64.197	12	5.350	0.939	0.949
Pair Cancellation	64.318	12	5.360	0.936	0.947	50.736	12	4.228	0.951	0.961
Picture Recall	60.248	12	5.021	0.940	0.951	38.922	12	3.244	0.961	0.972
Planning	61.173	12	5.098	0.939	0.949	37.184	12	3.099	0.963	0.974
Retrieval Fluency	65.401	12	5.450	0.940	0.950	66.227	12	5.519	0.940	0.950
Rapid Picture Naming	66.924	12	5.577	0.934	0.945	51.21	12	4.268	0.951	0.962
Sound Blending	95.94	12	7.995	0.918	0.926	78.812	12	6.568	0.930	0.939
Spatial Relations	59.397	12	4.950	0.942	0.952	42.434	12	3.536	0.958	0.969
Visual-Auditory Learning	59.667	12	4.972	0.945	0.955	69.604	12	5.800	0.939	0.948
Visual Matching	80.249	12	6.687	0.924	0.934	51.726	12	4.311	0.950	0.960
Verbal Comprehension	74.671	12	6.223	0.953	0.960	57.844	12	4.820	0.966	0.973

Table C.1b

Phase 2 Receptive Language Fit Indices for Ages 7 to 10

	Test					Cross-Validation				
	$\chi 2$	df	$\chi 2$/df	NFI	CFI	$\chi 2$	df	$\chi 2$/df	NFI	CFI
Analysis Synthesis	87.509	12	7.292	0.93	0.938	68.341	12	5.695	0.941	0.950
Auditory Attention	70.321	12	5.860	0.933	0.943	69.645	12	5.804	0.934	0.944
Auditory Working Memory	76.016	12	6.335	0.933	0.942	59.379	12	4.948	0.945	0.955
Concept Formation	77.276	12	6.440	0.949	0.956	65.673	12	5.473	0.957	0.964
Decision Speed	79.082	12	6.590	0.931	0.940	65.434	12	5.453	0.940	0.950
Delayed Recall Visual-Auditory Learning	67.016	12	5.585	0.940	0.95	57.882	12	4.824	0.948	0.958
General Information	104.642	12	8.720	0.931	0.938	107.037	12	8.920	0.923	0.930
Incomplete Words	62.933	12	5.244	0.944	0.953	77.158	12	6.430	0.927	0.937
Memory for Words	63.347	12	5.279	0.942	0.952	63.389	12	5.282	0.941	0.951
Numbers Reversed	74.219	12	6.185	0.935	0.944	73.845	12	6.154	0.933	0.942
Pair Cancellation	79.013	12	6.584	0.933	0.942	63.957	12	5.330	0.945	0.954
Picture Recall	63.504	12	5.292	0.939	0.949	41.739	12	3.478	0.958	0.969
Planning	72.415	12	6.035	0.930	0.940	59.330	12	4.944	0.942	0.952
Retrieval Fluency	66.846	12	5.571	0.942	0.951	62.819	12	5.235	0.942	0.952
Rapid Picture Naming	66.258	12	5.522	0.936	0.946	57.530	12	4.794	0.944	0.955
Sound Blending	90.799	12	7.567	0.923	0.931	88.476	12	7.373	0.923	0.932
Spatial Relations	69.17	12	5.764	0.936	0.946	63.215	12	5.268	0.939	0.949
Visual-Auditory Learning	63.644	12	5.304	0.946	0.955	63.928	12	5.327	0.945	0.954
Visual Matching	102.208	12	8.517	0.913	0.922	74.014	12	6.168	0.934	0.943
Verbal Comprehension	72.905	12	6.075	0.946	0.954	67.197	12	5.600	0.951	0.959

Table C.2a

Phase 2 Expressive Language Fit Indices for Ages 11 to 14

	Test					Cross-Validation				
	$\chi2$	df	$\chi2$/df	NFI	CFI	$\chi2$	df	$\chi2$/df	NFI	CFI
Analysis Synthesis	76.927	12	6.411	0.929	0.938	81.125	12	6.760	0.941	0.949
Auditory Attention	54.840	12	4.570	0.947	0.957	41.881	12	3.490	0.969	0.977
Auditory Working Memory	67.190	12	5.599	0.937	0.947	67.988	12	5.666	0.951	0.959
Concept Formation	78.232	12	6.519	0.932	0.941	72.806	12	6.067	0.949	0.957
Decision Speed	62.807	12	5.234	0.940	0.950	48.174	12	4.015	0.963	0.972
Delayed Recall Visual-Auditory Learning	77.826	12	6.486	0.928	0.938	65.098	12	5.425	0.952	0.960
General Information	81.192	12	6.766	0.943	0.951	72.291	12	6.024	0.960	0.966
Incomplete Words	58.530	12	4.878	0.945	0.955	81.509	12	6.792	0.942	0.950
Memory for Words	68.906	12	5.742	0.936	0.945	49.491	12	4.124	0.965	0.973
Numbers Reversed	80.053	12	6.671	0.925	0.934	57.679	12	4.807	0.957	0.965
Pair Cancellation	65.314	12	5.443	0.937	0.947	39.853	12	3.321	0.997	0.978
Picture Recall	51.081	12	4.257	0.950	0.961	41.882	12	3.490	0.968	0.976
Planning	76.670	12	6.389	0.927	0.937	50.947	12	4.245	0.961	0.969
Retrieval Fluency	88.066	12	7.339	0.921	0.930	64.955	12	5.412	0.952	0.960
Rapid Picture Naming	62.855	12	5.238	0.940	0.950	49.577	12	4.131	0.962	0.970
Sound Blending	65.059	12	5.422	0.942	0.951	52.620	12	4.385	0.963	0.971
Spatial Relations	66.622	12	5.552	0.937	0.947	50.984	12	4.249	0.962	0.970
Visual-Auditory Learning	96.748	12	8.062	0.915	0.924	65.888	12	5.491	0.953	0.961
Visual Matching	57.586	12	4.799	0.944	0.955	51.875	12	4.323	0.960	0.969
Verbal Comprehension	73.425	12	6.119	0.954	0.961	75.919	12	6.327	0.960	0.966

Table C.2b

Phase 2 Receptive Language Fit Indices for Ages 11 to 14

	Test					Cross-Validation				
	$\chi2$	df	$\chi2$/df	NFI	CFI	$\chi2$	df	$\chi2$/df	NFI	CFI
Analysis Synthesis	40.194	12	3.350	0.964	0.974	54.160	12	4.513	0.964	0.972
Auditory Attention	40.209	12	3.351	0.961	0.972	47.541	12	3.962	0.966	0.974
Auditory Working Memory	40.721	12	3.393	0.963	0.973	50.220	12	4.185	0.964	0.972
Concept Formation	40.954	12	3.413	0.973	0.980	52.083	12	4.340	0.972	0.978
Decision Speed	46.635	12	3.886	0.957	0.967	58.140	12	4.845	0.959	0.966
Delayed Recall Visual-Auditory Learning	39.231	12	3.269	0.963	0.974	49.929	12	4.161	0.964	0.972
General Information	68.003	12	5.667	0.949	0.957	58.829	12	4.902	0.966	0.972
Incomplete Words	46.210	12	3.851	0.954	0.965	77.382	12	6.449	0.946	0.953
Memory for Words	48.565	12	4.047	0.955	0.965	48.709	12	4.059	0.966	0.974
Numbers Reversed	43.427	12	3.619	0.960	0.970	46.415	12	3.868	0.967	0.975
Pair Cancellation	54.049	12	4.504	0.951	0.961	54.479	12	4.540	0.960	0.968
Picture Recall	40.561	12	3.380	0.959	0.970	49.894	12	4.158	0.962	0.971
Planning	48.259	12	4.022	0.953	0.963	42.791	12	3.566	0.968	0.976
Retrieval Fluency	48.062	12	4.005	0.955	0.965	44.057	12	3.671	0.968	0.977
Rapid Picture Naming	50.939	12	4.245	0.949	0.960	39.538	12	3.295	0.970	0.979
Sound Blending	48.761	12	4.063	0.955	0.965	48.278	12	4.023	0.967	0.974
Spatial Relations	45.261	12	3.772	0.956	0.967	49.890	12	4.158	0.963	0.972
Visual-Auditory Learning	44.233	12	3.686	0.960	0.970	47.536	12	3.961	0.968	0.975
Visual Matching	45.398	12	3.783	0.958	0.969	59.760	12	4.980	0.958	0.966
Verbal Comprehension	69.463	12	5.789	0.948	0.956	53.587	12	4.466	0.968	0.975

Table C.3a

Phase 2 Expressive Language Fit Indices for Ages 15 to 18

	Test					Cross-Validation				
Test	χ2	df	χ2/df	NFI	CFI	χ2	df	χ2/df	NFI	CFI
Analysis Synthesis	73.420	12	6.118	0.932	0.941	110.160	12	9.180	0.877	0.887
Auditory Attention	60.570	12	5.048	0.940	0.951	72.385	12	6.032	0.915	0.926
Auditory Working Memory	58.453	12	4.871	0.943	0.953	90.178	12	7.515	0.895	0.906
Concept Formation	68.392	12	5.699	0.938	0.948	99.687	12	8.307	0.895	0.905
Decision Speed	61.490	12	5.124	0.938	0.949	0.905	12	0.075	0.905	0.916
Delayed Recall Visual-Auditory Learning	68.524	12	5.710	0.935	0.945	113.866	12	9.489	0.873	0.882
General Information	65.843	12	5.487	0.953	0.961	80.032	12	6.669	0.936	0.944
Incomplete Words	60.132	12	5.011	0.942	0.953	71.279	12	5.940	0.918	0.930
Memory for Words	62.523	12	5.210	0.940	0.951	83.086	12	6.924	0.905	0.916
Numbers Reversed	71.740	12	5.978	0.930	0.940	84.989	12	7.082	0.902	0.913
Pair Cancellation	70.914	12	5.910	0.929	0.939	70.667	12	5.889	0.913	0.925
Picture Recall	73.365	12	6.114	0.927	0.938	75.489	12	6.291	0.908	0.92
Planning	59.967	12	4.997	0.939	0.950	62.213	12	5.184	0.923	0.936
Retrieval Fluency	82.138	12	6.845	0.922	0.932	96.898	12	8.075	0.892	0.903
Rapid Picture Naming	61.852	12	5.154	0.938	0.949	71.803	12	5.984	0.914	0.926
Sound Blending	58.018	12	4.835	0.946	0.956	79.371	12	6.614	0.917	0.928
Spatial Relations	64.873	12	5.406	0.936	0.946	68.131	12	5.678	0.920	0.932
Visual-Auditory Learning	70.524	12	5.877	0.935	0.944	101.480	12	8.457	0.891	0.901
Visual Matching	74.155	12	6.180	0.926	0.936	75.959	12	6.330	0.908	0.920
Verbal Comprehension	68.150	12	5.679	0.953	0.961	83.925	12	6.994	0.938	0.946

Table C.3b

Phase 2 Receptive Language Fit Indices for Ages 15 to 18

	Test					Cross-Validation				
	$\chi2$	df	$\chi2$/df	NFI	CFI	$\chi2$	df	$\chi2$/df	NFI	CFI
Analysis Synthesis	26.103	12	2.175	0.977	0.988	66.081	12	5.507	0.936	0.946
Auditory Attention	25.895	12	2.158	0.974	0.986	52.982	12	4.415	0.941	0.953
Auditory Working Memory	26.329	12	2.194	0.975	0.986	50.252	12	4.188	0.945	0.957
Concept Formation	30.566	12	2.547	0.979	0.987	55.856	12	4.655	0.956	0.965
Decision Speed	30.650	12	2.554	0.972	0.983	54.496	12	4.541	0.943	0.954
Delayed Recall Visual-Auditory Learning	29.619	12	2.468	0.972	0.983	67.810	12	5.651	0.930	0.941
General Information	53.181	12	4.432	0.958	0.967	68.064	12	5.672	0.945	0.953
Incomplete Words	30.987	12	2.582	0.97	0.981	54.987	12	4.582	0.939	0.951
Memory for Words	22.578	12	1.882	0.979	0.990	56.090	12	4.674	0.941	0.952
Numbers Reversed	28.274	12	2.356	0.974	0.985	50.650	12	4.221	0.949	0.960
Pair Cancellation	39.654	12	3.305	0.964	0.974	55.162	12	4.597	0.94	0.951
Picture Recall	33.840	12	2.820	0.966	0.977	46.424	12	3.869	0.946	0.959
Planning	31.050	12	2.588	0.969	0.980	50.129	12	4.177	0.943	0.955
Retrieval Fluency	25.907	12	2.159	0.976	0.987	53.569	12	4.464	0.943	0.954
Rapid Picture Naming	26.107	12	2.176	0.974	0.985	51.832	12	4.319	0.941	0.953
Sound Blending	30.904	12	2.575	0.971	0.982	66.297	12	5.525	0.933	0.943
Spatial Relations	23.961	12	1.997	0.977	0.988	52.588	12	4.382	0.941	0.953
Visual-Auditory Learning	27.566	12	2.297	0.975	0.986	52.171	12	4.348	0.947	0.958
Visual Matching	39.624	12	3.302	0.963	0.974	55.042	12	4.587	0.941	0.953
Verbal Comprehension	40.479	12	3.373	0.968	0.977	61.424	12	5.119	0.949	0.958

Appendix D – Phase 2 Path Coefficients

Table D.1a

Phase 2 Expressive Language Path Coefficients for Ages 7 to 10

	Test		Cross-Validation	
	Cultural Loading → Expressive Language	Expressive Language → Test	Cultural Loading → Expressive Language	Expressive Language → Test
Analysis Synthesis	0.30	0.49	0.32	0.35
Auditory Attention	0.32	0.24	0.32	0.22
Auditory Working Memory	0.30	0.45	0.33	0.39
Concept Formation	0.27	0.57	0.28	0.50
Decision Speed	0.31	0.22	0.31	0.14
Delayed Recall Visual-Auditory Learning	0.31	0.42	0.30	0.58
General Information	0.38	0.95	0.39	0.89
Incomplete Words	0.37	0.48	0.35	0.29
Memory for Words	0.32	0.37	0.32	0.35
Numbers Reversed	0.30	0.31	0.31	0.18
Pair Cancellation	0.31	0.21	0.31	0.20
Picture Recall	0.31	0.19	0.26	0.01
Planning	0.30	0.16	0.29	0.16
Retrieval Fluency	0.28	0.50	0.28	0.51
Rapid Picture Naming	0.31	0.33	0.35	0.38
Sound Blending	0.32	0.51	0.35	0.45
Spatial Relations	0.31	0.29	0.29	0.14
Visual-Auditory Learning	0.32	0.50	0.25	0.58
Visual Matching	0.29	0.32	0.31	0.17
Verbal Comprehension	0.24	0.98	0.26	1.01

Table D.1b

Phase 2 Receptive Language Path Coefficients for Ages 7 to 10

	Test		Cross-Validation	
	Cultural Loading → Expressive Language	Expressive Language → Test	Cultural Loading → Expressive Language	Expressive Language → Test
Analysis Synthesis	0.19	0.77	0.07	0.63
Auditory Attention	0.23	0.38	0.10	0.40
Auditory Working Memory	0.21	0.65	0.14	0.55
Concept Formation	0.16	0.87	0.06	0.85
Decision Speed	0.15	0.50	0.01	0.42
Delayed Recall Visual-Auditory Learning	0.24	0.67	0.16	0.67
General Information	0.38	0.92	0.34	0.86
Incomplete Words	0.34	0.51	0.25	0.32
Memory for Words	0.29	0.46	0.13	0.44
Numbers Reversed	0.21	0.55	0.05	0.48
Pair Cancellation	0.14	0.51	0.00	0.43
Picture Recall	0.28	0.31	0.17	0.10
Planning	0.28	0.27	0.21	0.33
Retrieval Fluency	0.24	0.60	0.18	0.48
Rapid Picture Naming	0.29	0.37	0.21	0.26
Sound Blending	0.29	0.54	0.24	0.52
Spatial Relations	0.22	0.45	0.03	0.35
Visual-Auditory Learning	0.26	0.67	0.13	0.64
Visual Matching	0.16	0.50	0.03	0.46
Verbal Comprehension	0.24	0.93	0.25	0.96

Table D.2a

Phase 2 Expressive Language Path Coefficients for Ages 11 to 14

	Test		Cross-Validation	
	Cultural Loading → Expressive Language	Expressive Language → Test	Cultural Loading → Expressive Language	Expressive Language → Test
Analysis Synthesis	0.37	0.41	0.38	0.41
Auditory Attention	0.36	0.18	0.36	0.32
Auditory Working Memory	0.38	0.42	0.38	0.50
Concept Formation	0.35	0.53	0.35	0.56
Decision Speed	0.32	0.17	0.31	0.13
Delayed Recall Visual-Auditory Learning	0.37	0.42	0.38	0.40
General Information	0.32	0.92	0.36	0.94
Incomplete Words	0.38	0.33	0.45	0.51
Memory for Words	0.34	0.26	0.39	0.50
Numbers Reversed	0.37	0.26	0.37	0.28
Pair Cancellation	0.30	0.14	0.28	0.11
Picture Recall	0.28	0.15	0.24	0.08
Planning	0.32	0.15	0.34	0.16
Retrieval Fluency	0.34	0.51	0.39	0.36
Rapid Picture Naming	0.37	0.33	0.30	0.07
Sound Blending	0.38	0.46	0.39	0.49
Spatial Relations	0.35	0.24	0.36	0.26
Visual-Auditory Learning	0.28	0.56	0.37	0.54
Visual Matching	0.31	0.14	0.31	0.10
Verbal Comprehension	0.31	1.01	0.31	0.99

Table D.2b

Phase 2 Receptive Language Path Coefficients for Ages 11 to 14

	Test		Cross-Validation	
	Cultural Loading → Expressive Language	Expressive Language → Test	Cultural Loading → Expressive Language	Expressive Language → Test
Analysis Synthesis	0.06	0.68	0.18	0.68
Auditory Attention	0.13	0.46	0.24	0.49
Auditory Working Memory	0.11	0.73	0.23	0.65
Concept Formation	0.06	0.85	0.16	0.83
Decision Speed	0.03	0.41	0.15	0.42
Delayed Recall Visual-Auditory Learning	0.06	0.59	0.22	0.57
General Information	0.24	0.93	0.36	0.91
Incomplete Words	0.17	0.38	0.36	0.52
Memory for Words	0.08	0.51	0.29	0.58
Numbers Reversed	0.04	0.55	0.22	0.55
Pair Cancellation	0.03	0.38	0.21	0.38
Picture Recall	0.06	0.30	0.28	0.23
Planning	0.04	0.35	0.27	0.34
Retrieval Fluency	0.11	0.55	0.26	0.51
Rapid Picture Naming	0.12	0.35	0.29	0.25
Sound Blending	0.15	0.55	0.26	0.57
Spatial Relations	0.05	0.40	0.24	0.42
Visual-Auditory Learning	0.05	0.64	0.23	0.61
Visual Matching	0.03	0.42	0.20	0.46
Verbal Comprehension	0.23	0.97	0.32	0.94

Table D.3a

Phase 2 Expressive Language Path Coefficients for Ages 15 to 18

	Test		Cross-Validation	
	Cultural Loading → Expressive Language	Expressive Language → Test	Cultural Loading → Expressive Language	Expressive Language → Test
Analysis Synthesis	0.42	0.61	0.33	0.51
Auditory Attention	0.45	0.33	0.39	0.31
Auditory Working Memory	0.46	0.45	0.39	0.48
Concept Formation	0.42	0.60	0.29	0.60
Decision Speed	0.41	0.21	0.37	0.20
Delayed Recall Visual-Auditory Learning	0.44	0.54	0.32	0.52
General Information	0.41	0.92	0.36	0.94
Incomplete Words	0.47	0.47	0.41	0.45
Memory for Words	0.45	0.43	0.39	0.37
Numbers Reversed	0.45	0.36	0.39	0.39
Pair Cancellation	0.43	0.16	0.29	0.03
Picture Recall	0.46	0.31	0.35	0.15
Planning	0.41	0.17	0.31	0.11
Retrieval Fluency	0.42	0.47	0.34	0.53
Rapid Picture Naming	0.46	0.34	0.38	0.31
Sound Blending	0.44	0.49	0.39	0.55
Spatial Relations	0.45	0.31	0.35	0.28
Visual-Auditory Learning	0.43	0.54	0.28	0.60
Visual Matching	0.44	0.18	0.35	0.14
Verbal Comprehension	0.38	0.97	0.29	1.03

Table D.3b

Phase 2 Receptive Language Path Coefficients for Ages 15 to 18

	Test		Cross-Validation	
	Cultural Loading → Expressive Language	Expressive Language → Test	Cultural Loading → Expressive Language	Expressive Language → Test
Analysis Synthesis	0.20	0.69	0.11	0.71
Auditory Attention	0.21	0.39	0.18	0.41
Auditory Working Memory	0.22	0.55	0.19	0.60
Concept Formation	0.20	0.83	0.11	0.81
Decision Speed	0.17	0.47	0.09	0.45
Delayed Recall Visual-Auditory Learning	0.21	0.56	0.13	0.60
General Information	0.32	0.84	0.31	0.89
Incomplete Words	0.27	0.51	0.23	0.44
Memory for Words	0.24	0.59	0.16	0.49
Numbers Reversed	0.20	0.52	0.17	0.61
Pair Cancellation	0.18	0.48	0.10	0.39
Picture Recall	0.24	0.36	0.19	0.32
Planning	0.20	0.32	0.15	0.41
Retrieval Fluency	0.23	0.55	0.21	0.53
Rapid Picture Naming	0.23	0.40	0.21	0.40
Sound Blending	0.25	0.55	0.22	0.57
Spatial Relations	0.21	0.47	0.15	0.40
Visual-Auditory Learning	0.20	0.57	0.15	0.64
Visual Matching	0.18	0.47	0.10	0.43
Verbal Comprehension	0.29	0.88	0.26	0.94