2024

EVALUATION OF CULTURAL-LINGUISTIC FACTORS IN SPEECH-LANGUAGE TESTING WITH ENGLISH LEARNERS

Mayra Alejandra Reyes Ruiz

EVALUATION OF CULTURAL-LINGUISTIC FACTORS IN SPEECH-LANGUAGE TESTING WITH ENGLISH LEARNERS

A dissertation submitted in partial fulfillment
of the requirements for the degree of

DOCTOR OF PSYCHOLOGY

to the faculty of the

DEPARTMENT OF PSYCHOLOGY

of

ST. JOHN'S COLLEGE OF LIBERAL ARTS AND SCIENCES

at

ST. JOHN'S UNIVERSITY
New York
by
Mayra Alejandra Reyes Ruiz

Date Submitted 4/19/2024 Date Approved 4/19/2024

Mayra Alejandra Reyes Ruiz Samuel O. Ortiz, Ph. D.

ABSTRACT

EVALUATION OF CULTURAL-LINGUISTIC FACTORS IN SPEECH-LANGUAGE TESTING WITH ENGLISH LEARNERS

Mayra Alejandra Reyes Ruiz

As the population of schools increases in diversity, school psychologists are challenged to provide equitable assessments for English Language Learners. However, due to lack of training in non-discriminatory assessment, inconsistent administration practices, and lack of knowledge of language development, in some districts, ELLs are overrepresented in Special Education, including Speech Language Impairment. For fair and a valid assessment, it is imperative that practitioners understand the cultural and linguistic development factors that affect ELL test performance. The Culture-Language Test Classifications, and Culture-Language Interpretive Matrix provide practitioners with information on the degree of cultural loading and linguistic demand required on a test, and provide a tool for interpretation of validity of results to determine the likely presence of disability. Prior research on test performance of ELLs has yielded patterns of expected test performance based on the categories set by the C-LIM based on the expected attenuation of performance due to the degree cultural and linguistic demand. The purpose of this study is to compare the performance of ELLs determined eligible for Speech and Language Impairment to the expected performance to determine whether the obtained test performance is consistent to the predictions of the C-LIM classifications. Results indicate that ELL performance on tests that make up speech-language evaluations, decrease as a function of increasing cultural and linguistic development required in English. When compared to the normative mean, ELLs perform significantly lower across all levels, and effect sizes increase as the C-LIM levels increase which account for the combined effect of culture and language. When compared to the expected performance of ELLs eligible for SLI with a moderate difference in language development and acculturative experiences there is a lack of significant differences suggesting test performance is as expected by the C-LIM. Additionally, a subtest was identified to have higher language demand than previously classified, and recommendations were made for re-classification to a higher level on the C-LTC and C-LIM. Results from this study highlight the need for complete comprehensive assessment of cultural and linguistic factors to adequately categorize degree of language difference for appropriate test interpretation.

ACKNOWLEDGEMENTS

This dissertation project would not have been possible without the support and contributions of many along this journey. I would like to express my gratitude to my advisor Dr. Samuel Ortiz for his generosity of expertise, wisdom and guidance as we navigated this research. I am greatly appreciative of my committee members Dr. Mark Terjesen and Dr. Marlene Sotelo-Dynega for their support, and contributions through the completion of this endeavor, that allowed me to persevere to this point.

I am also grateful to all my clinical supervisors at the SJU clinic, and all those that poured into me in all externships and internship, for inspiring me to believe in the value I bring to this work. It is with their confidence in me that I have developed as a clinician and professional. All children and families I've had the privilege to serve in training, especially those from culturally and linguistically diverse backgrounds, continuously ignited my sense of purpose and desire to complete this project and degree.

Lastly, I'd like to acknowledge my dear peers, friends and family that made the journey attainable. Whether co-working in person or from afar, employing many pomodoros, and virtual tools, I am grateful to my peers and colleagues that have joined efforts to reach this end. I have my sister, brothers, family and friends to thank for healing laughter and reenergizing spirits. Finally, I was propelled by my mother's validating and encouraging reminder that it is difficult to create something that hasn't existed yet, and accomplishing that, in itself, is a special endeavor. I am grateful to her for all her sacrifices that made this all possible for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Over the last several decades the population in the US has been shifting toward a majority composed of ethnic, cultural, and linguistic minorities (Rhodes et al., 2005). As school psychologists adhering to ethical practice standards for diverse student populations, the responsibility of nondiscriminatory assessment of bilingual students falls on all practitioners, bilingual or monolingual. These changes in the composition of the public-school classroom pose ethical obligations for equitable access to education and nondiscriminatory assessment as part of an eligibility determination for special education services.

Diana vs State Board of Education (1970) is a landmark case of a Mexican American student who was placed in classes for the mentally retarded based on the invalid use of intelligence tests scores to diagnose mental retardation (Rhodes et al., 2005). Larry P. v. Riles (1979) is another seminal case in which an African American student was tested with culturally loaded intelligence tests that yielded an invalid score of his intellectual abilities, thus, intelligence tests are not allowed in the education evaluations of African American students in California in an effort to eliminate invalid and incorrect identification of an intellectual disability. The Individuals with Disability Education of 2004, (IDEA; PL 94-142, 2004) aims to prevent the inappropriate overidentification or disproportionate representation by race and ethnicity of children as children with disabilities including children with disabilities with a particular impairment. School psychologists carry an ethical and legal responsibility to be mindful of ways to prevent discriminatory practices especially in assessment and evaluations that determine special education placements.

Population of Minorities in Schools

Enrollment of minority groups in public elementary and secondary schools is projected to increase, particularly that of Hispanic students by 8 percent, Black students by 1 percent, and Asian/Pacific Islander students by 20 percent between 2016 and 2028, while enrollment of students who are white are expected to decrease by 7 percent. Further, enrollment of students who identify with two or more races is expected to increase by 51 percent in the same timeframe (NCES, 2021). Nationally, the percentage of English Language Learners (ELLs) in public schools is already increasing from 9.2 in 2010 (4.5 million students) to 10.2 percent in 2018 (5.0 million students) (NCES, 2021).

Of all English Language Learners, Hispanic students constitute the majority with 3.8 million students enrolled (77.6 percent) followed by Asians with 528,700 students (10.7 percent), White ELLs (6.7 percent), and Black ELLs with 218,000 students (4.4 percent) (NCES, 2019). Fewer than 40,000 students were identified as ELL in the remaining racial and ethnic groups for which data were collected (Pacific Islanders, American Indians/Alaska Natives, and individuals of two or more races). In 2018, 15.3 percent of the ELL population were identified as having a learning disability (National Center for Education Statistics, 2021, Irwin et al., 2024). Despite the changing composition of the public-school population, the educational system continues to struggle in providing adequate and equitable special education services for culturally and linguistically diverse students with learning difficulties (Sullivan, 2011). This growing subgroup of students poses a challenge for a school psychologist tasked with the delivery of equitable practices to prevent disproportionate representation of children with a particular impairment by race and ethnicity (IDEA, 2004).

LITERATURE REVIEW

Disproportionate Identification of English Language Learners in Special Education

Several decades of research have elucidated consistent patterns of disproportionate representation of racial and ethnic minorities in special education with some variation among racial/ethnic groups, disability categories, and location samples (Artiles, 2005; Artiles & Kilinger, 2006; Artiles et al., 2010; Dunn, 1968; Skiba et al., 2008; Sullivan, 2011; Sullivan & Bal, 2013;). Findings consistently show that African American students are overrepresented in overall special education categories in nearly all states, twice as likely to be classified with an Emotional Disturbance, and 2.7 times as likely to be classified as Intellectual Disability. Meanwhile, American Indian/Alaska Native students are twice as likely to be classified with a Specific Learning Disability. In some districts, ELLs are overrepresented at twice the rate of their White peers in the educational categories of Mild Intellectual Disability, Specific Learning Disability, and Speech-Language Impairment. Conversely, Latinos and Asian/Pacific Islanders tend to be underrepresented across categories in national samples (Sullivan & Bal, 2013; U.S Dept. of Education, 2010). Overrepresentation has been theorized to be more pronounced when minority students represent a larger proportion of the population, while there are underrepresentation patterns in national studies (Skiba et al., 2008).

There is ample evidence in disproportionality research indicating that groups who are misclassified and disproportionately represented in special education are negatively affected by factors such as stigmatization, lowered expectations, fewer opportunities to learn, substandard instruction, and isolation from the general education environment (Albrecht et al, 2011). These implications of disproportional overidentification result in further discriminatory practices that propel and/or perpetuate institutional racism. That is by misclassifying minority students, they are more likely to have poor life outcomes as a result of institutional practices of discriminatory assessment (Artiles et al., 2010; De Valenzuela, 2006; Sullivan et al., 2008; Sullivan, 2009, 2011; Sullivan & Bal, 2013).

Disproportionality in special education results from a complex interaction of factors including practitioner cultural competency in assessment including methodological and conceptual issues (Rhodes et al., 2005). Although research has acknowledged the existence of psychometric test biases, current measures do not account for these biases and are largely still not equipped to account for them. Some of the factors that contribute to the disproportionality in special education are the test selection and interpretation practices associated with identifying culturally and linguistically diverse children in special education. Professionals untrained in nondiscriminatory assessment can mistakenly interpret low cognitive verbal scores and language scores as a sign of a potential disability (Nieves-Brull, 2006).

Other variables include practitioner knowledge of new language acquisition processes understanding of the bilingual brain, benefits of native-language instruction, and discerning between normal language development and disordered language (Dollaghan 2007; Dollaghana et al., 2011). A typically developing child learning English as a new language will develop conversational proficiency or Basic Interpersonal Communication Skills (BICS) within one or three years of English exposure. BICS are less cognitively demanding and highly contextualized. Conversely, an ELL will develop Cognitive Academic Language Proficiency which is grade-appropriate, more complex, in five to seven years (Cummins, 1984; Rhodes et al., 2005; Ortiz, 2014). Educators and practitioners often mistake a proficiency in social conversation as a similar proficiency in academic contexts comparable to their monolingual peers. This assumption can lead practitioners to test cognitive or language abilities in English though CALP is not yet achieved and obtaining lower scores that are incorrectly interpreted as an impairment (Rhodes et al., 2005). It is important to recognize that many cognitive and language assessments evaluate advanced cognitive academic language abilities (CALP) and are highly language-loaded. Culturally and linguistically diverse students tend to score lower than native monolingual English speakers, but this outcome does not necessarily indicate a speech impairment or language disorder. These issues contribute to the disproportionate identification of English Language Learners in the special education classification of Speech-Language Impairment in many states.

Assessment of Language for SLI

Commonly selected norm-referenced language tests as reported by school-based speech-language pathologists in their assessments of speech-language disorders are the Clinical Evaluation of Language Fundamentals (CELF), Peabody Picture Vocabulary Test (PPVT), Preschool Language Scale (PLS), Expressive One-Word Picture Vocabulary Tests (EOWPVT), Comprehensive Assessment of Spoken Language (CASL), Oral and Written Language Scales (OWLS) (Betz et al., 2013; Caesar & Kohler, 2009; Ogiela & Montzka, 2021).

The Clinical Evaluation of Language Fundamentals, 5th edition (CELF; Semel et al., 2013) is a norm-referenced battery of tests designed to assess a variety of language skills in children and young adults (ages 5-21). The 14 tests within the CELF battery measure receptive and expressive skills, reading, writing, pragmatics, and memory. The

Core Language Score is a composite score of the Receptive Language Index, Expressive Language Index, Language Content Index, Language Structure Index, and Language Memory Index. The CELF-V was standardized on a normative sample similar to the 2012 U.S. Census data for age, sex, race/ethnicity, geographic region, and caregiver education. The sample included 2,380 students ages 5 to 21 and 56.9% of the sample was White, 20% Hispanic, 13.8% African American, and 3.6% Asian (Hutchins & Pratt, 2017). The developers report that the participants "spoke and understood English very well" (Technical Manual, p.32).

The Peabody Picture Vocabulary Test, 5th edition (PPVT-5; Dunn, 2019) is a measure of receptive vocabulary for individuals ages 2 years 6 months to 90 years of age and older. It provides an assessment of word knowledge including nouns, verbs, and attributes. The PPVT-5 was standardized on a stratified sample representative of U.S Census estimates in 2017 and included only English speakers. The fifth edition incorporates a digital administration requiring an individual to select one out of four corresponding images after listening to a verbal prompt of the target word. The utility of the PPVT-5 is as a screening tool for the identification of receptive language disorders and supplemental narrow assessment of vocabulary knowledge which is an important component of speech and language (Canivez & Graham Laughlin, 2021).

The Pre-school Language Scales, 5th edition (PLS-5; Zimmerman et al., 2011) is a measure of language development designed for use with children from birth through 7 years 11 months. The test assesses receptive and expressive language skills in the Auditory Comprehension, Expressive Communication, and Total Language scale scores. It is used to identify children in preschools and early intervention programs with language delays, and receptive, expressive, and mixed-language disorders. The normative sample was stratified by the 2008 U.S Census data. Development of this 5th edition included bias review by a panel to determine whether items were appropriate for children from various ethnic groups, SES groups, and regions of the US. Items were eliminated using an Item Response Theory bias analysis (McKnight & Shapley, 2014).

The Comprehensive Assessment of Spoken Language, 2nd Edition is a measure of oral language for individuals 3 to 21 years of age. The test is designed to be an in-depth assessment of skills for the purpose of identifying language delays and disorders, oral language skills strengths and weaknesses, and English proficiency in English Language Learners. The 14 tests of the CELF-2 battery provide a General Language Ability Index along with a Receptive Language Index, Expressive Language Index, Lexical/Semantic Index, Supralinguistic Index). The CASL-2 was developed with a normative sample that reflects the 2012 U.S Census data. This test includes scoring guidance for speakers of African American English or similar dialect, but no mention of considerations for Latino or Asian influence in English syntax (Moyle & Newman, 2021).

The Expressive One-Word Vocabulary Test, 4th edition (EOWPVT, Brownell, 2011) is a measure of total acquired expressive vocabulary skills, for individuals aged 2 to 80 or older. The test assesses an individual's ability to name objects, actions, and concepts when presented with single illustrations. The EOWPVT manual indicates that the test can aid in the documentation of vocabulary development, supplementing reading assessments, screening for early, language delay, and indirect cognitive skills. The EPWPVT provides only a total score of vocabulary. The norming sample included

English-speaking individuals from only 26 states of the U.S, and the manual states that results should be interpreted with caution if the individuals do not match the description of the normative sample (Harris & Johnson, 2014).

The Ortiz Picture Vocabulary Acquisition Test (Ortiz PVAT; Ortiz, 2018) measures receptive vocabulary acquisition in native English speakers and English Learners ages 2 years 6 months to 22 years 11 months. It is a digitally administered assessment of English vocabulary knowledge and instructions are available in 5 languages. Target words were developed with consideration of the progression of Basic Interpersonal Communicative Skills (BICS) and Cognitive Academic Language Proficiency (CALP). The Ortiz PVAT has two distinct sets of norms, a monolingual native English speaker norming group and an English Learner norms adjusted for lifetime exposure to English. Examiners consider the age of first exposure to English, learning experiences, formal education, parental proficiency, and SES, along with other factors to yield a percent lifetime exposure to English. This feature allows for administration in English with individuals learning English, regardless of heritage language, and does not require bilingual examiners. Individual's performance is compared to others of the same age with the same percentage of lifetime exposure for a more fair and valid normative comparison. The Ortiz PVAT can be used (Alfonso et al., 2020; Barrett & Matthews, 2021; Tello, 2020).

The Bilingual English-Spanish Assessment (BESA; Peña et. al., 2018) is a measure of language development created by speech-language pathologists. This assessment considers 17 Spanish dialects and uses language exposure norms to better evaluate students with limited exposure to English. However, this assessment may be administered only to bilingual individuals whose heritage language is Spanish, and who are transitioning from preschool into an elementary school setting (Alfonso et al., 2020).

Many of these tests were normed on a population sample that is representative of the U.S Census data available at the time. The sample is typically stratified with considerations for race, ethnicity, and SES, but these cannot be equated with the representation of cultural and linguistic experiences. Tests are inherently biased, as they are linguistically and culturally loaded reflecting the norms, ideology, and language of the population they represent. Research has indicated that those with language backgrounds that vary significantly from the individuals comprising the norm, will tend to score lower on standardized tests as a function of the language demands (Flanagan et al., 2013). Fair and valid evaluation of English learners goes beyond testing. Evaluations must be in alignment with the principles of IDEA, that is students must be evaluated in a way that is nondiscriminatory and equitable.

Non-discriminatory Assessment Rationale and Process

Given the changing school-age populations in racial, ethnic, cultural, and linguistic diversity, school-based practitioners have been faced with evaluating students who are culturally and linguistically different from them, often speaking an unfamiliar language. Consequently, school psychologists and speech-pathologists have developed unstandardized methods of evaluating culturally and linguistically diverse students (Ortiz, 2014). Surveyed school psychologists, endorse the selection of nonverbal tests, dominant language testing, and modifications to the administration of cognitive tests including translation, (Ochoa, 2004; Sotelo-Dynega & Dixon, 2014). The use of native language testing may still result in discriminatory practice as bilingual, bicultural students are compared to monolingual single-culture norm groups. These tests are limited for use by culturally competent and linguistically fluent evaluators (Ortiz, 2014). Training of school psychologists often falls short in providing training in biased or discriminatory assessment practices and how to avoid them, and specific training in non-discriminatory assessment framework is lacking. (Harris et al., 2021)

In light of these limiting and unstandardized practices, Ortiz (2002) developed a 10-step nondiscriminatory assessment framework to systematically address biases inherent to the data collection and interpretation processes when assessing ELL students' cognitive and academic functioning. The hypothesis-driven approach serves to guide equitable decisions, minimize bias, and the framework seeks to uncover relevant information and data in a systematic manner. The prereferral process, which features eight steps, is structured to avoid unnecessary evaluations and reduce the risk of misclassification by assessing and evaluating : (a) the purpose for the intervention, (b) using authentic and alternative procedures that include work samples, criterion-referenced assessments, and dynamic assessments, (c) the learning ecology (i.e., extrinsic factors in the learning environment that may impede learning) (d) the language proficiency in their native language and English (i.e., BICS and CALP), (e) the opportunity for learning (i.e. appropriate level of instruction), (f) the opportunity for learning (i.e., appropriate level and setting of instruction), (g) the educationally relevant cultural and language factors (i.e., level of acculturation). the evaluation and revision of hypotheses, and re-testing hypotheses, (i) the determination of the need for an evaluation, and the appropriate language(s) of formal assessment. Pre-referral procedures of data collection may lead to a comprehensive evaluation of a disability. In the ninth step of the framework, the aim is to *(j)reduce potential bias in the traditional assessment practices* while considering that bias exists even in native language testing, and awareness of the norming population. Finally, in the final step of the post-referral procedures practitioners should (k) support conclusions via data convergence and multiple indicators to make meaning of the data and the impact of culture and language (Ortiz, 2014).

Culture - Language Interpretive Matrix

Concerns about the validity of test performance and obtained scores are essential in the evaluation of culturally and linguistically diverse individuals as the main goal is to determine whether scores are a reflection of a potential disorder or adversely impacted by culture and language resulting in a difference rather than a disorder. To make any meaningful interpretations from collected assessment data, especially in high-stakes decision-making situations, it is critical to ascertain the extent to which the validity of the test scores may have been compromised (Flanagan et al., 2013).

The Culture-Language Test Classifications (C-LTC; Flanagan et al., 2000; Flanagan & Ortiz, 2001; Flanagan, et al., 2007; McGrew & Flanagan, 1998) was developed as an extension of the CHC theoretical classifications, to identify tests with low cultural loading and linguistic demand that may generate more valid scores in evaluations for English Language Learners. The Culture-Language Test Classifications (C-LTC) and the Culture-Language Interpretation Matrix (C-LIM) were developed in attempts to identify the tests with most validity when measuring the cognitive abilities of the McGrew Gf-Gc and Catell-Horn-Carroll theory (McGrew & Flanagan, 1998; Flanagan & Ortiz, 2001). The goal of the C-LTC was to identify the tests with the lowest levels of cultural loading and language demand as these would likely yield the most valid scores. The C-LTC sorts tests into three classification levels (low, moderate, high) on both dimensions of cultural loading and linguistic demands. These two dimensions are based on evidence in the research on performance patterns of culturally and linguistically diverse (CLD) students on cognitive and achievement tests in English, and the factors that could invalidate test results (Flanagan et al., 2013). Standardized norm-referenced tests are organized by three classification levels (low, moderate, high) based on results of empirical research, expert consensus, and examination of task characteristics. Test scores are assigned low, moderate, or high cultural loading, and language demand classifications based on their deviation from the normative mean (e.g SS=100, Scs=10). Those yielding scores near the normative mean are classified as low cultural and low linguistic demand. Tests with scores that are more than one standard deviation away from the mean (e.g SS=85, Scs=7) are classified as high cultural loading and high linguistic demand and are most attenuated by cultural and linguistic influence. Scores between these two points are considered moderate classification. Altogether these three classifications on two dimensions result in 5 levels of classifications on a 3x3 matrix (Flanagan et al., 2013). Scores that decrease from left to right indicate attenuation test performance by language, and a decrease from top to bottom indicate attenuation based on cultural difference. Scores that decrease along the diagonal from upper left to bottom right cells indicate the combined effects of culture and language difference on test performance. I general, (Flanagan et al., 2013; Sotelo-Dynega et al., 2011).

The Culture-Language Interpretive Matrix (C-LIM; Flanagan & Ortiz, 2001; Flanagan, Ortiz & Alfonso, 2007; Ortiz, 2001, 2004; Ortiz & Flanagan, 1998) was designed to assist in interpretation by evaluating whether results obtained from standardized testing are either valid (permitting interpretation) or due primarily to influence of cultural or linguistic variables (precluding interpretation) (Ortiz et al., 2018). The C-LIM along with the C-LTC are systematic nondiagnostic approaches to address the validity of obtained test results based on empirical research (Flanagan et al., 2013). Test performance of a CLD individual may be attenuated by the correlated effects of culture and language such that a decline in scores is apparent and expected across the matrix from left to right. If the pattern of scores entered into the matrix reveals a systematic decline across the cells, the scores can be interpreted as invalid measures of the construct because they are primarily influenced by cultural loading and linguistic demands, not the intended test construct. If there is an absence of a systematic decline in scores as linguistic and cultural demands increase, then the scores can be interpreted as a valid measure. Even without a clear decline, cultural and linguistic differences may be present, but in this case, they are not the primary influence on the obtained scores Flanagan et al., 2013).

The interaction between the two dimensions on the C-LTC yields nine cells which are further organized into levels based on equivalent cultural and linguistic attenuation. Tests in the top left cell (i.e. Level 1) are expected to produce the highest scores as the cell contains subtests with low cultural loading and low language demand. Tests in the bottom right (i.e., Level 5) are expected to produce the lowest scores as the cell contains subtests with high cultural loading and high language demand. The diagonals represent equivalent degree of attenuation and make up the remaining 3 levels. These are Low Culture/Moderate Language and Moderate Culture/Low Language (i.e., Level 2), the main diagonal with Low Culture/High Language, Moderate Culture/Moderate Language and High Culture/Low Language (i.e., Level 3), and the High Culture/Moderate Language and Moderate Culture/High Language (i.e., Level 4) (Flanagan et al., 2013). The C-LTC provides a categorical system for understanding the impact cultural and linguistic factors have on bilingual student test performance (Flanagan et al., 2013). Based on research findings of bilingual test performance, and expert consensus, subtests are assigned into one of nine cells based on the two dimensions and this systematic categorization aids in understanding the impact of cultural and linguistic factors on test performance.

Due to variation in acculturation, instructional language, language proficiency, and socioeconomic factors, there are differences among students designated as bilinguals or English Language Learners. These differences are categorized based on degree of difference on a range from slightly different, moderately different, and markedly different (Flanagan et al., 2013). Obtained scores deviating from the normative mean by one standard deviation or more are greatly influenced by culture and language, SS=85 (Level 5, high/high). Scores between these endpoint cells for typically developing ELLs for other cells are SS=89 (Level 4, high/moderate), SS=92 (Level 3, low/high), SS=95 (Level 2, moderate/low). The expected mean scores on the C-LIM for ELLs who are moderately different are Level 1= 94, Level 2=91, Level 3=87, Level 4=82, and Level 5=77 (Dynda, 2008).

Prior research using the C-LIM for ELL Assessment

Most studies of the Cognitive-Language Interpretive Matrix as a validity evaluation tool focus on cognitive and achievement tests (Aguera Verderosa, 2007; Brown, 2008; Calderón-Tena et al., 2020; Dhaniram-Beharry, 2008, 2008; Meyer, 2013;

Nieves-Brull, 2006; Ortiz et al., 2018; Sotelo-Dynega et al., 2013; Styck & Watkins, 2013; Yoo, 2015). Studies have shown that the C-LIM (Flanagan & Ortiz, 2001) is a viable method for discerning the difference between performance on a standardized assessment due to a cultural and linguistic difference (typical pattern of decline) from a student with an educational disability (no systematic pattern of decline) based on obtained test scores (Flanagan et al., 2007).

Often in research, bilingual individuals are compared to monolingual English-speaking individuals, however bilingual or English Language Learners are not a homogenous group, and there are significant variations in language development, and acculturating experiences among them. Researchers have identified three degrees of difference; slightly, moderate, and markedly different (Dynda, 2008; Flanagan & Ortiz, 2001; Sotelo-Dynega et al., 2013). These categories are determined based on linguistic factors such as levels of English proficiency and literacy, and acculturative factors such as duration of U.S residence, and parental education. The expected performance on testing with English Language Learners is attenuated by cultural and linguistic factors in an increasing fashion as C-LIM Levels increase. For example, students determined to be moderately different are expected to score less than the normative mean by 5-7 points on Level 1, 7-10 points on Level 2, 10-15 points on Level 3, 15-20 points on Level 4, and 20-30 points less on Level 5 (Flanagan et al., 2013) .

A limited number of studies on the use of C-LIM as a method for analyzing patterns of scores for CLD individuals evaluated for Speech-Language Impairments are available. In validation studies for use in SLI evaluations, findings generally support the distinct performance patterns of CLD school-age and preschool students with speech/

language problems and global delays (Lella Souravlis, 2010; Vasquez, 2010). Tychanska (2009) found a pattern of decline that suggests that cultural and linguistic factors did affect test scores, and English learners with SLI had a steeper pattern of decline than English speakers with SLI. With an established pattern of decline in a sample of SLI, further research is needed to validate the current classification of the C-LIM for language tests commonly used in SLI evaluations.

PURPOSE AND HYPOTHESES

School Psychologists are in a unique position to provide nondiscriminatory assessments, equitable and fair evaluations to all students including English language learners from ethnic and linguistic minorities. Culturally competent practitioners regardless of bilingualism seek to select valid assessments with appropriate norming and can interpret the cultural and language impact that may differentiate a speech-language disorder above and beyond a language difference. Though some research validating the C-LIM for use as a viable method to distinguish between a student demonstrating low performance on a standardized assessment due to cultural and linguistic differences from a student with an educational disability, most of these focus on cognitive and achievement assessments. Some research suggests a likely marked increase in attenuation on tests of language ability (Lella Souravlis, 2010; Tychanska, 2009). While some research exists on distinguishing a speech-language impairment from a language difference, there is a need for more research on the use of the C-LIM to assist such interpretations.

The purpose of this study is to replicate prior research outcomes of patterns of attenuation of ELL performance found on cognitive testing compared to the normative sample to demonstrate impact of increasing cultural loading and linguistic demands on test scores with speech-language measures. Secondly the study aimed to provide support for the usage of linguistic and cultural factors, in consideration of degree of language difference in order to aid in determining validity of test scores when assessing for a speech-language disability. The clinical significance of this study is to increase awareness of the culture loading and linguistic demands of speech and language assessments commonly used to determine Speech-Language Impairment eligibility and examine the utility of C-LIM in evaluations with English Language Learners.

Hypotheses

Based on prior research and existing theories in literview of assessment of English Language Learners, the following were the hypotheses of this study.

1. It was hypothesized that English Language Learners referred for evaluations, would score significantly lower than the normative group mean.
2. It was hypothesized that English Language Learners referred for Speech Language Disorder would perform in a manner that is consistent with moderately different English Language Learners in consideration of language proficiency and language differences.

METHOD

Participants

Standardized testing and demographic data were retrieved from an ongoing study utilizing student special education records from public schools in Central California. The information collected from the files were multidisciplinary reports, psychoeducational reports, and speech and language reports. From these reports student age, grade, sex, ethnicity, state English proficiency scores (ELPAC), English Language Learner (ELL) status, cognitive testing scores, academic achievement scores, and speech/language testing scores, and special education eligibility. The sample for this study included 119 participants with a mean age of 8.02, ranging from 4 to 12, and a mean grade of 2.72, ranging from kindergarten to sixth grade. All 119 evaluation reports available were for bilingual English-Spanish speaking English Language learners. English Language Learner status was determined based on report of a home language different from English in the social history, report of their placement in an integrated or pull-out English Language Learner class, or report of their ELPAC scores for English Proficiency (N=57). The ELPAC score mean (M=1.171) for 57 students provided data on students with a low or developing English Language Proficiency, and the remainder of the students were determined to have average or higher language proficiency as they are not referred for ELL services or proficiency testing. Though individually, some students would be markedly different, and others would be slightly different based on their English proficiency and accompanying cultural experiences, the group as a whole, is best represented as moderately different.

Special education classification eligibility was determined via eligibility statements in evaluation reports. Students in the archival sample were mostly found eligible for SLD (N=103), followed by SLI (N=20), and OHI (N=10). Students eligible for OHI were also determined to be eligible for SLD. There were 17 students identified as eligible for both SLD and SLI per their evaluation reports. Just three students did not qualify for special education services. While students with a Specific Learning Disability were not a primary target of this study, their data was included in analysis as patterns of performance for LD have been established in the existing research and provide a comparison group for emerging patterns in this sample.

Criteria for Test Inclusion

Special education records were reviewed for students referred for special education evaluations. This study aimed to include tests administered in the determination of a speech language impairment for English Learners. To include assessment test scores, the study required standardized testing subtest scores from cognitive, and speech language testing. Subtests were included if at least 20 student evaluations reported a score. Tests that did not have an identified classification in the Culture-Language Test Classification (C-LTC), such as achievement testing, were excluded. Though there were a variety of cognitive measures administered including bilingual and Spanish language measures (e.g., Bateria-Cog, C-TONI, CAS-2, DAS-2, KABC-2, RIAS, TONI-4, UNIT-2, WJ-Cog, WNV, WPPSI-4), and speech and language measures (e.g.,CASL-2, CELF-4, CELF-5, EOWPVT, EVT-3, FAR, GFTA-3, OWLS-2, PLS-5, PPVT-5, ROWPVT, SPELT, TAPS-3, TAPS-3Span, TOLD-5) they were not administered frequently across test batteries such that they could be analyzed. Cognitive and Speech-Language standardized tests that met the inclusion criteria were the five subtests of the Weschler Intelligence Scales for Children- fifth edition, the 7 subtests of the Comprehensive Test of Phonological Processing (CTOPP-2), and the 7 subtests of the Language Processing Skills Assessment (TAPS-4).

Instruments

Wechsler Intelligence Scale for Children- 5th edition (WISC-V)

The Wechsler Intelligence Scale for Children- 5^{th} edition is a measure of cognitive abilities and is administered to children ages six through sixteen years old (Weschler, 2014). The WISC-V contains 10 primary tests. The ten primary subtests of the WISC-V are Similarities, Vocabulary, Block Design, Matrix Reasoning, Figure Weights, Digit Span, Picture Span, Coding and Symbol Search. These primary subtests measure five cognitive domains: Verbal Comprehension, Visual Spatial, Working Memory, Fluid Reasoning, and Processing Speed. As part of the standardization process, the WISC-V was normed on a nationally representative sample of 2,200 children, closely matched to the 2012 U.S Census on demographics of race/ethnicity, parent education, and gender balanced (Weschler, 2014). At the time of standardization, special group studies were performed with participants with known characteristics such as disability (e.g., ADHD, Autism, Intellectual Disability, Specific Learning Disability), and English Language Learners. In the special groups study, there were 16 English Language Learners, of which 50% were females, and 88% were Hispanic. ELLs were compared to match control participants from the norm sample. This study found that ELLs scored significantly lower than the control participants on Verbal Comprehension, Working Memory, and the overall FSIQ. Notably, there were no significant differences between groups on subtests requiring minimal expressive and receptive language (Weschler, 2014)

CTOPP-2 Comprehensive Test of Phonological Processing, 2nd Edition (CTOPP-2)

The Comprehensive Test of Phonological Processing, 2nd Edition is measure of phonological processing skills for individuals ages 4:0 through 24, 11 months. The test is composed of twelve subtests: Elision, Blending Words, Sound Matching (4-6), Phoneme Isolation, Blending Words, Segmenting Non-words, Memory for Digits, Nonword Repetition, Rapid Digit Naming, Rapid Letting Naming, Rapid Color Naming (4-6), and Rapid Object Naming (4-6). Subtest scores have a mean of 10 and standard deviation of 3. The subtests are then organized into component and composite standard scores: Phonological Awareness, Phonological Memory, Rapid Symbolic Naming, and Rapid Non-symbolic Naming. Related to the psychometric properties, the CTOPP-2 reports strong internal consistency with an average coefficient exceeding .80 for all subtests except Nonword Repetition with an average alpha of .77 (Wagner, et al., 2013). The test-retest correlations for the subtests ranged from .75 to .92, and inter-rater reliability coefficients surpassed .90. Validity of the CTOPP-2 relative to measures with related constructs resulted in average coefficients ranging from .49 to .84 for the subtests. The CTOPP-2 was standardized with a norming group of 1,990 individuals aged 4 to 24. The sample was reported as representative of the 2010 U.S Census, based on gender, ethnicity, Hispanic status, family income, and education attainment of parents (Wagner et al., 2013).

Language Processing Skills Assessment, 4th Edition (TAPS-4)

The is a measure of language processing and comprehension skills for individuals ages five through 21. The test is composed of 11 subtests: Word Discrimination, Phonological Deletion, Phonological Blending, Syllabic Blending (*sup*), Number Memory Forward, Word Memory, Sentence Memory, Number Memory Reversed (*sup*), Processing Oral Direction, Auditory Comprehension, Auditory Figure-Ground (*sup*). These subtests yield three composite scores: Phonological Processing, Auditory Memory, and Listening Comprehension. These domains are important for the development of listening, communication, and higher order language skills (Martin et al., 2018). The TAPS-4 normative sample were 2023 individuals aged 5 through 21 that were matched to the U.S population demographic characteristics with in a 5% match. The normative sample included individuals with specific language impairment, learning disability, auditory processing disorder, ADHD, and hearing impairment. In terms of psychometric properties, the TAPS-4 technical manual reports internal consistency coefficients above .80, and test-retest reliability coefficients above. 93 for the index scores. Interrater reliability coefficients were high, above .97, for processing oral directions, the auditory figure-ground, auditory comprehension subtests, and the listening comprehension index. In terms of validity, the CELF-5 Receptive Language Index (Clinical Evaluation of Language Fundamentals, 5th edition), correlated positively (.68-.80) with the Processing Oral Directions, Auditory Comprehension, and Listening Comprehension scores of the TAPS-4 (Martin et al., 2018).

Culture-Language Test Classifications & Interpretive Matrix

The Culture-Language Test Classifications (C-LTC) categorizes subtests by degree of cultural loading and linguistic demand on a scale of low, moderate or high. The two dimensions are represented on a 3 by 3 matrix with degree of cultural loading on the left axis and degree of linguistic demand on the top axis (Flanagan et al., 2013). The Culture-Language Interpretive Matrix (C-LIM) is a tool to evaluate concerns of score validity when testing bilingual individuals with English measures in order to determine a language or learning difference versus a disorder. The C-LIM represents obtained subtest score data in the same C-LTC nine-cell table, converted to a standard score (M=100, SD=15). All the subtest data in each cell is aggregated to achieve means for each C-LIM Cell. To compare trends and patterns of performance, the Cell data is then grouped by Level of combined cultural loading and linguistic demand. Subtests in the top left cell (i.e. Cell 1, Level 1) are expected to produce the highest scores as the cell contains subtests with low cultural loading and low language demand. The three diagonals represent equivalent degree of attenuation. These are Low Culture/Moderate Language and Moderate Culture/Low Language (i.e., Cells 2 and 4, Level 2), the main diagonal with Low Culture/High Language, Moderate Culture/Moderate Language and High Culture/Low Language (i.e., Cells 3, 5, 7, Level 3), and the High Culture/Moderate Language and Moderate Culture/High Language (i.e., Cells 6 and 8, Level 4). Tests in the bottom right (i.e., Cell 9, Level 5) are expected to produce the lowest scores as the cell contains subtests with high cultural loading and high language demand. (Flanagan et al., 2013). The degree of language proficiency and acculturation further impacts the expected performance of English Language Learners, and in this sample they are determined to be

Moderately different based on language proficiency testing available. English Language Learners that are moderately different from the mainstream norming group are expected to score less than the normative mean by 5-7 points on Level 1, 7-10 points on Level 2, 10-15 points on Level 3, 15-20 points on Level 4, and 20-25 points less on Level 5 (Flanagan et al., 2013). Further, established comparison values for moderately different English Language Learners were Level 1= 94, Level 2=91, Level 3 = 87, Level 4 = 82, and Level 5=77 (Dynda, 2008; Sotelo-Dynega et al., 2013) were used in this analysis .

Procedure

All available student files from archived from another study were reviewed and coded for relevant variables. Data collected from the participant records were, age, grade, gender, ethnicity, ESL status, English proficiency level and test scores. No identifying information were gathered. The overall sample consisted of 119 participants that were referred for initial or triennial evaluations for special education. Students were tested with various cognitive, achievement, speech and language measures. Subtest scores were converted to standard scores (M=100, SD=15) for ease of comparison. Descriptive analyses were conducted for frequency of use of each subtest, and those with an N of 20 or more were included in subsequent analyses. Given this criteria, only the subtests from the Weschler Intelligence Scales for Children, fifth edition (WISC-V), the Comprehensive Test of Phonological Processing, second edition (CTOPP-2) and the Language Processing Skills Assessment (TAPS-4) were used consistently throughout the sample and had a sufficient frequency for meaningful analysis. The ten subtests from the WISC-V and the seven subtests from the CTOPP-2, and the seven subtests of the TAPS-4 were sorted into the C-LIM based on established C-LTC classification.

Statistical Analyses

All obtained tests scores from the evaluations were analyzed for frequency of use in the sample and subtests were selected for further analysis based on inclusion criteria of a frequency of 20 scores. The data that met inclusion criteria were test scores from the Weschler Intelligence Scales for Children, fifth edition (WISC-V), the Comprehensive Test of Phonological Processing, second edition (CTOPP-2) and the Language Processing Skills Assessment (TAPS-4). Based on the C-LTC and C-LIM, the ten core subtests of the WISC-V are classified into cells 1,2,3,5, and 9. The CTOPP-2 has seven subtests classified into cells 2, 3, 5, 6 and 9. The TAPS-4 has seven subtests classified into cells 3, 5, 6, and 9. There were no test scores classified into C-LIM cells 4, 7 and 8.

Means for the 24 included subtests were categorized based on their C-LTC classification into the nine cells of the Culture-Language Interpretive Matrix. Variables for each C-LIM Cell were computed. Subtest scale scores were converted to a standard score, and means were calculated for each of the six cells that had tests scores classified. Then C-LIM level means were obtained by combining subtest scores according to established arrangement of degree of cultural loading and linguistic demand. The Level 1 composite included subtests in Cell 1, Level 2 included those in Cell 2 and 4, Level 3 included those in Cells 3, 5, and 7, Level 4 included those in Cells 6 and 8, and Level 5 included data in Cell 9. The 24 subtests included in this study did not have subtests categorized in cells 4, 7 or 8.

Once data was organized by C-LIM Cells and Levels, the data was further organized by special education eligibility classifications. Two main groups were created for comparisons, students eligible for Speech Language Impairment, and those eligible for Specific Learning Disability. Cases identified as eligible for Speech Language Impairment were grouped into separate variables for each Cell and each Level. Variables for Specific Learning Disability were created in the same fashion. All analyses were conduced by full sample, and also by disability eligibility groups (e.g, SLD, SLI).

One-sample t-test comparing obtained scores by referred ELLs to the normative mean by cell

To examine differences between the obtained scores from ELL student evaluations and the normative mean of 100, one-sample test was conducted with the whole sample by C-LIM Cell. There were six one-sample t-tests conducted by cell with the overall sample, as 3 cells did not have data. The results of this analysis established presence of a difference between obtained scores from a referred sample of ELLs and the norm group in lieu of a monolingual English speaking student sample. This statistical analysis procedure was repeated with data grouped by disability classification. A one-sample t-test was conducted for each SLI Cell and each SLD Cell comparing obtained scores to the normative mean.

One-sample t-test comparing the obtained scores by referred ELLs to the normative mean by level

To further evaluate patterns of performances and difference from the normative group, composite variables were computed to organize C-LIM cell means into C-LIM Level means. The means from cell 1 remained in Level 1, the data in cells 2 and 4 (no data) were combined for Level 2, data in cells 3, 5, and 7 (no data) were combined for Level 3, data in cells 6 and 8 (no data) were combined for Level 4, and data in cell 9 is equivalent to Level 5. All but Level 3 are composed of just one cell. Each of the Levels were compared to the normative mean with one-sample t-tests to evaluate differences in performance by level representing the combined impact of cultural loading and linguistic demand. This statistical analysis procedure was repeated with data grouped by disability classification. A one-sample t-test was conducted for each SLI Level and each SLD Level comparing obtained scores to the normative mean.

One-sample t-test comparing obtained scores for moderately different ELLs to expected means by Levels

To examine the difference between the sample of ELLS referred for evaluation, and the expected pattern of performance of moderately different ELLs, additional one-sample t-tests were conducted. The procedures of analysis were repeated by comparison of mean of each C-LIM Level to the expected mean for an ELL with a moderate difference in cultural and linguistic experiences. Whereas the first analysis compared each level and all cells to the normative mean of 100, in this analysis each one-sample t-test for the five levels contained a different comparison value. The comparison values by level as determined by Dynda (2008) for a moderately different English Language Learner were Level 1= 94, Level 2=91, Level 3 = 87, Level 4 = 82, and Level 5=77. One sample t-tests were conducted for each Level for the overall referral group, the SLI and SLD group, comparing obtained scores to those expected for a ELL student with moderate language differences.

One-sample t-test comparing expected mean for moderately different ELLs to obtained scores by Cells

To determine differences in performance based on cultural loading and language demand, further analyses were conducted by C-LIM Cells. One sample t-tests were conducted for each cell with data compared to the expected performance of an ELL with a moderate language difference. Each cell's comparison value corresponded to the values expected of their Levels as identified above. One-sample t-tests were conducted to determine differences in performance of the overall referred sample, and groups of those that qualified for SLI and SLD.

RESULTS

The purpose of this study is to compare the performance of English Language Learners with Speech and Language Impairment to the performance of monolingual English speakers to determine a pattern of performance is consistent to the predictions of the C-LIM classifications. Patterns of performance have been established for monolingual English speakers without a disability (similar to the normative sample), English Learners without a disability (systematic decline as a function of increasing culture and language, English Learners with Learning Disabilities (random pattern, limited deficits, and somewhat declining). Student participants in this study were all students who were considered as bilingual, and under half were administered the ELPAC, indicating their English Language Proficiency is developing. The data gathered did not include students who only speak English, so for the purposes of group comparisons, the normative mean, with similarity to a monolingual English speaker is used for analysis and evaluation of the study's hypothesis about individuals referred for a Speech Language Impairment.

Descriptive Statistics

The overall sample consisted of archival 119 cases of evaluations with 58.8 percent male students and 41.2 percent female students (SD = .49). The mean age of the participants was 8.03 years (SD=1.72) ranging from 4 to 12 years old. At the time of evaluation, students were in grades Kindergarten through sixth grade, with a mean grade of 2.72 (SD=1.64) as shown in Table 1. Students in the sample were referred for initial (N=75) and triennial (N=42) evaluations for special education. The overall sample included cases that were determined to be eligible for Speech Language Disability

(N=20), Specific Learning Disability (N=103), Other Health Impairment (N=10), or both SLD and SLI (N=17). All students in the sample were identified as English Language Learners, having some Hispanic heritage, being in ELL programming, and some in dual language classes. About half of ELL students (N=57) were administered the English Language Proficiency Assessment for California (ELPAC) which is scored on a range of zero to four increasing in English Proficiency. The mean ELPAC score was 1.71 (SD=.80) across the sample administered the state test, and frequency of results are seen in Table 2. Ethnicity and primary Language was recorded according to their Psychoeducational Evaluations. For this sample all students were Hispanic, and Bilingual Spanish/English speakers. Table 1 provides a summary of these demographic descriptive statistics and Table 2 provides a frequency view of demographic characteristics by overall sample and by disability classification.

Table 1

Summary Statistics for Overall Sample

Demographic Variable	N	M	SD	Range
Gender	119	.41	0.50	0-1
Age	119	8.03	1.72	4-12
Grade	119	2.66	1.64	0-6
ELPAC	57	1.71	0.80	0-4

Table 2

Demographic Characteristics

Characteristic	SLI		SLD		Full sample	
	n	%	*n*	%	*n*	%
Gender						
Female	10	50.0	43	41.7	49	41.2
Male	10	50.0	60	58.3	70	58.8
Age						
4	-	-	1	1.0	1	0.80
5	1	5.0	1	1.0	3	2.50

6	3	15.0	17	16.50	17	14.30
7	9	45.0	32	31.10	34	28.60
8	2	10.0	21	20.40	21	17.60
9	2	10.0	14	13.60	16	13.40
10	3	15.0	11	10.70	15	12.60
11	-	-	6	5.80	9	7.60
12	-	-	-	-	3	2.50
Grade						
Kindergarten	1	5.0	1	1.0	3	2.5
1st	9	45.0	33	32.0	35	29.4
2nd	4	20.0	26	25.2	26	21.8
3rd	2	10.0	20	19.4	21	17.6
4th	3	15.0	10	9.7	11	9.2
5th	1	5.0	10	9.7	16	13.4
6th	-	-	3	2.9	7	5.9
Evaluation						
Initial	9	45.0	70	68.0	75	63.0
Reevaluation	11	55.0	31	30.1	42	35.3
Re-eval not tested	-	-	2	1.9	2	1.7
ELPAC						
0	-	-	1	2.0	1	1.8
1	6	54.50	21	41.2	24	42.1
2	3	27.30	21	41.2	23	40.4
3	1	9.10	7	13.7	8	14.0
4	1	9.10	1	2.0	1	1.8

Note. There were 103 cases identified as SLD, and 20 cases identified as SLI. Of these17 cases are dual classified as both SLD and SLI.

Test data were obtained from the Weschler Intelligence Scale for Children, Fifth Edition (WISC-V), The Comprehensive Test of Phonological Processing, Second Edition (CTOPP-2), and the Language Processing Skills Assessment, Fourth Edition (TAPS-4). The subtests from these assessments met inclusion criteria of N greater than 20. Notably, the most frequently administered test in the sample of 119 were Elision (N=59) and Blending Words (N=59) from the CTOPP-2. The least frequently administered of the selected tests is the Phonological Deletion subtest (N=20) of the TAPS-4. Table 3 represents the frequency of each subtest of all three included tests, and the original scale score along with the transformed standard score. These standard scores were then categorized into the C-LIM cells using previously determined cultural loading and linguistic demand levels in the C-LTC. Further analyses use the resulting means of the nine C-LIM cells and C-LIM levels (Table 4) of cultural loading and linguistic demand. Though each C-LIM Level is composed of 1-3 cells, some levels only contain data in one cell. C-LIM Cell means are in Table 5. In fact, in this dataset the only C-LIM Level with more than one cell filled with data is Level 3, which is comprised of cells three and five. As a result, data in Level 1 is equivalent to Cell 1, Level 2 is equivalent to Cell 2, Level 3 is the mean of Cells 3 and 5, Level 4 is equivalent to Cell 6, and Level 5 is equivalent to Cell 9.

Table 3

Performance of ELL Referred Students on WISC-V, CTOPP-2, and TAPS-4

		Scale Scores		Standard Scores	
	n	M	SD	M	SD
WISC-V					
Similarities	38	6.18	3.00	80.92	14.97
Vocabulary	38	6.84	2.52	84.21	12.60
Block Design	39	8.44	2.26	92.18	11.29
Visual Puzzles	41	8.85	2.57	94.27	12.87
Matrix Reasoning	41	8.12	3.01	90.61	15.05
Figure Weights	41	8.85	2.60	94.27	13.02
Digit Span	40	5.55	2.28	77.75	11.38
Picture Span	41	7.46	2.25	87.32	11.24
Coding	40	7.90	3.02	89.50	15.10
Symbol Search	37	8.46	2.85	92.30	14.27
CTOPP-2					
Elision	59	5.25	2.07	76.27	10.36
Blending Words	59	6.12	2.44	80.51	12.24
Phoneme Isolation	44	5.68	1.74	78.41	8.679
Memory for Digits	50	5.66	2.16	78.30	10.81
Nonword Repetition	51	6.35	3.26	81.96	16.10
Rapid Digit Naming	52	7.13	2.24	85.67	11.20
Rapid Letter Naming	47	6.77	2.43	83.83	12.17
TAPS-4					
Word Discrimination	21	7.86	2.18	89.29	10.87
Phonological Deletion	20	5.50	2.69	77.50	13.43

Phonological Blending	21	6.81	2.60	84.05	13.00
Number Memory Forward	24	7.21	2.36	86.04	11.79
Word Memory	24	6.42	3.06	82.08	15.32
Sentence Memory	24	6.17	2.50	80.83	12.48
Processing Oral Directions	21	6.05	2.20	80.24	11.01

Table 4

Summary of Mean Scores on Five C-LIM Levels

	N	Range	M	SD
Level 1 – Cell 1	41	72-120	92.44	12.27
Level 2 – Cells 2, 4	76	55-108	83.50	11.17
Level 3 – Cells 3, 5, 7	90	58-107	84.22	8.54
Level 4 – Cells 6, 8	82	62-108	79.69	9.10
Level 5 – Cell 9	78	55-100	79.80	10.07

Table 5

Summary of Mean Scores on Nine C-LIM Cells

		N	M-Ex	M- Ob	SD	Range
Cell 1	LL/LC	41	94	92.44	12.27	72 - 120
	Visual Puzzles			94.27	12.87	
	Matrix Reasoning			90.61	15.05	
Cell 2	ML/LC	76	91	83.50	11.17	55- 108
	Memory for Digits			78.30	10.81	
	Block Design			92.18	11.29	
	Picture Span			87.32	11.24	
	Coding			89.50	15.10	
	Symbol Search			92.30	14.27	
Cell 3	HL/LC	82	87	83.15	10.51	55 - 115
	Number Memory Forward			86.04	11.79	
	Rapid Digit Naming			85.67	11.20	
	Digit Span			77.75	11.38	
Cell 4	LL/MC	-	91	-	-	-
Cell 5	ML/MC	86	87	86.24	14.60	55-120
	Word Discrimination			89.29	10.87	
	Word Memory			82.08	15.32	
	Nonword Repetition			81.96	16.10	
	Figure Weights			94.27	13.02	

Cell 6	HL/MC	82	82	79.69	9.10	62-108
	Phonological Deletion			77.50	13.43	
	Phonological Blending			84.05	13.00	
	Sentence Memory			80.83	12.48	
	Elision			76.27	10.36	
	Blending Words			80.51	12.24	
	Rapid Letter Naming			83.83	12.17	
Cell 7	LL/HC	-	87	-	-	-
Cell 8	ML/HC	-	82	-	-	-
Cell 9	HL/HC	78	77	79.80	10.07	55-100
	Processing Oral Direction			80.24	11.01	
	Phoneme Isolation			78.41	8.68	
	Similarities			80.92	14.97	
	Vocabulary			84.21	12.60	

Note. No subtests in cells 4, 7, and 8.

Evaluation of Hypotheses

To evaluate whether the data collected on the performance of English Language Learners with Speech Language Impairment is consistent with prior research and reflective of a true disability, as well as reflective of the classifications suggested by the C-LIM a series of one-sample t-tests were conducted along with an analysis of effect size with Cohen's *d*. Effect size measured by Cohen's *d* values operationally defined as a small effect (.2), moderate effect (.5) and large effect (.8) (Cohen, 1977). A standard score was calculated for each subtest with at least 20 scores reported. Subtests were organized into the C-LIM 9-cell matrix based on the previously determined classifications of cultural loading and language demand for each subtest. Average scores were calculated for each of the nine cells of the C-LIM. There were means for six of the nine cells of the matrix, and these mean scores were evaluated using on-sample t-tests to examine differences between them and the normative mean, and the predicted performance of a moderately different English Language Learner. The intent of this analysis is to evaluate whether the performance of English Language Learners referred for a suspected disability of Speech Language Impairment follows the pattern established in prior research.

Evaluation of Hypothesis 1

The first hypothesis in the present study predicts that English Language Learners referred for Special Education evaluation would perform significantly lower than monolingual English students, and in lieu of this group, would perform lower than the normative sample. Means were obtained for the ELL sample at each of the five C-LIM levels of cultural loading and linguistic demand and compared to the normative mean at each level.

Results of one-sample t-tests indicated that comparison of the five C-LIM Levels of cultural and language loading to the normative mean were statistically significantly different with p=<.001, and effect sizes ranging from moderate (-.62) to large (-2.23). The results of the overall sample is listed in Table 6 and by disability group on Table 7 and 8. The mean value of Level 1 scores (M=92.44, SD=12.27) was significantly lower than the normative mean performance on Level 1 tests; t(40)= -3.95, p=<.001. The effect size *d* of -.62 indicates a moderate effect. The mean value of Level 2 scores (M=83.50, SD=11.17) was significantly lower than the normative mean performance on Level 2 tests; t(75)= -12.88, p= <.001. The effect size *d* of -1.48 indicates a large effect. The mean value of Level 3 scores (M=84.22, SD=8.57) was significantly lower than the normative mean performance on Level 3 tests; t(89)= -17.54, p=<.001. The effect size *d* of -1.85 indicates a large effect. The mean value of Level 4 scores (M=79.70, SD=9.10) was significantly lower than the normative mean performance on Level 4 tests; t(81)= -20.20, p=<.001. The effect size *d* of -2.23 indicates a large effect. The mean value of Level 5 scores (M=79.80, SD=10.07) was significantly lower than the normative mean performance (M=100, SD= 15) on Level 5 tests; t(77)= -17.72, p=<.001. The effect size *d* of -2.01 indicates a large effect.

Table 6

Comparison of Overall Sample to Normative Mean by C-LIM Level

C-LIM Level	n	m	SD	*t*	*df*	*p*	*Cohen's d*
Level 1 - LL/LC	41	92.44	12.27	-3.95	40	<.001***	-.62
Level 2 - low/mod	76	83.50	11.17	-12.88	75	<.001***	-1.48
Level 3 - low/high	90	84.22	8.57	-17.54	89	<.001***	-1.85
Level 4 - mod/high	82	79.70	9.10	-20.20	81	<.001***	-2.23
Level 5 - HC/LC	78	79.80	10.07	-17.72	77	<.001***	-2.01

Note. Normative Mean (M=100), $*p < .05$. $**p < .01$. $***p < .001$

When comparing bilingual student performance classified with SLD to the normative mean, all C-LIM Level t-scores indicate a statistically significant difference, with a effect size ranging from moderate to large (-.58 to 2.54). The mean value of SLD Level 1 scores (M=92.84, SD=12.41) was significantly lower than the normative mean performance on Level 1 tests; t(36)=-3.51, p= .001. The effect size *d* of -.58 indicates a moderate effect. The mean value of SLD Level 2 scores (M= 83.54, SD=11.51) was significantly lower than the normative mean performance on Level 2 tests; t(68)= -11.886, p= <.001. The effect size *d* of -1.43 indicates a large effect. The mean value of SLD Level 3 scores (M= 84.25, SD= 8.77) was significantly lower than the normative mean performance on Level 3 tests; t(81)= -16.26, p= <.001. The effect size *d* of -1.80 indicates a large effect. The mean value of SLD Level 4 scores (M= 79.85, SD= 9.36) was significantly lower than the normative mean performance on Level 4 tests; t(75)= -

18.78, p= <.001. The effect size *d* of -2.15 indicates a large effect. The mean value of SLD Level 5 scores (M= 79.60, SD=10.30) was significantly lower than the normative mean performance on Level 5 tests; t(70)= -16.69, p= <.001. The effect size *d* of -1.97 indicates a large effect size.

Table 7

Comparison of SLD Group to Normative Mean by C-LIM Level

C-LIM Level	n	m	SD	*t*	*df*	*p*	*Cohen's d*
Level 1 - low/low	37	92.84	12.42	-3.51	36	.001 ***	-.58
Level 2 - low/mod	69	83.54	11.51	-11.89	68	<.001***	-1.43
Level 3 - low/high	82	84.25	8.77	-16.26	81	<.001***	-1.80
Level 4 - mod/high	76	79.85	9.36	-18.78	75	<.001***	-2.15
Level 5 - high/high	71	79.60	10.30	-16.69	70	<.001***	-1.97

Note. Normative Mean (M=100), **p* < .05. ***p* < .01. ****p* < .001

Looking at the results of the SLI group by C-CLIM Level, all levels were significantly different from the normative mean, and effect sizes are large across all levels, seen in Table 8. The mean value of SLI Level 1 scores (M= 90.56, SD=11.58) was significantly lower than the normative mean performance on Level 1 tests; t(8)= -2.45, p= .040. The effect size *d* of -.82 indicates a large effect. The mean value of SLI Level 2 scores (M= 86.91, SD=10.66) was significantly lower than the normative mean performance on Level 2 tests; t(12)= -4.427, p= <.001. The effect size *d* of -1.23 indicates a large effect. The mean value of SLI Level 3 scores (M= 82.41, SD=10.36) was significantly lower than the normative mean performance on Level 3 tests; t(13)= -6.36, p= <.001. The effect size *d* of -1.70 indicates a large effect. The mean value of SLI Level 4 scores (M= 76.96, SD= 9.06) was significantly lower than the normative mean performance on Level 4 tests; t(9)= -8.04, p= <.001. The effect size *d* of -2.54 indicates a large effect. The mean value of SLI Level 5 scores (M= 77.15, SD=15.85) was significantly lower than the normative mean performance on Level 5 tests; t(11)= -5.00, p= <.001. The effect size *d* of -1.44 indicates a large effect.

Table 8

Comparison of SLI Group to Normative Mean by C-LIM Level

C-LIM Level	n	m	SD	*t*	*df*	*p*	*Cohen's d*
Level 1 - low/low	9	90.56	11.58	-2.45	8	.04*	-0.82
Level 2 - low/mod	13	86.91	10.66	-4.43	12	<.001***	-1.23
Level 3 - low/high	14	82.41	10.36	-6.36	13	<.001***	-1.70
Level 4 - mod/high	10	76.96	9.06	-8.04	9	<.001***	-2.54
Level 5 - high/high	12	77.15	15.85	-5.00	11	<.001***	-1.44

Note. Normative Mean (M=100), **p* < .05. ***p* < .01. ****p* < .001

Results evaluating the difference between obtained scores of the overall sample by C-LIM cells and the normative mean indicate significant differences for all C-LIM Cells with available data at the <.001 level, and effect sizes ranging from moderate (-0.62) to large (-2.23). C-LIM Levels are composed of a combination of cells from the nine cell matrix, and since there is only data in Cells 1,2,3,5, and 6, and 9. All Levels are equivalent to the one cell with data available with the exception of Level 3, which is composed of Cells 3, and 5. In other words, in this data, Cell 1 is equivalent to Level 1 as usual, Cell 2 is equivalent to Level 2 in the absence of Cell 4, Cell 6 is equivalent to Level 4 in the absence of cell 8, and Cell 9 is equivalent to Level 5 as usual. Therefore the only C-LIM Level with further C-LIM cell analysis is Level 3 with Cells 3 and 5. The mean value of overall Cell 3 scores (M= 83.15, SD=10.51) was significantly lower than the normative mean performance on Level 1 tests; t(81)= -14.51, p= <.001. The effect size *d* of -1.60 indicates a large effect. The mean value of overall Cell 5 scores (M= 86.24, SD=14.60) was significantly lower than the normative mean performance on Level

1 tests; t(85)= -8.74, p= <.001. The effect size *d* of -0.94 indicates a large effect. These results are listed in Table 9.

Table 9

Comparison of Overall Sample to Normative Mean by C-LIM Cell

C-LIM Cell	n	m	SD	*t*	*df*	*p*	*Cohen's d*
Cell 1 - low/low	41	92.44	12.27	-3.95	40	<.001***	-0.62
Cell 2 - low/mod	76	83.50	11.17	-12.88	75	<.001***	-1.48
Cell 3 - low/high	82	83.15	10.51	-14.51	81	<.001***	-1.60
Cell 5 - mod/mod	86	86.24	14.60	-8.74	85	<.001***	-0.94
Cell 6 - mod/high	82	79.70	9.10	-20.20	81	<.001***	-2.23
Cell 9 - high/high	78	79.80	10.07	-17.72	77	<.001***	-2.01

Note. Normative Mean (M=100), *$p < .05$. **$p < .01$. ***$p < .001$

Table 10

Comparison of SLD Group to Normative Mean by C-LIM Cell

C-LIM Cell	n	m	SD	*t*	*df*	*p*	*Cohen's d*
Cell 1 - low/low	37	92.84	12.42	-3.51	36	.001***	-0.58
Cell 2 - low/mod	69	83.54	11.51	-11.89	68	<.001***	-1.43
Cell 3 - low/high	74	83.49	10.20	-13.92	73	<.001***	-1.62
Cell 5 - mod/mod	78	85.86	14.34	-8.71	77	<.001***	-0.99
Cell 6 - mod/high	76	79.85	9.36	-18.78	75	<.001***	-2.15
Cell 9 - high/high	71	79.64	10.31	-16.63	70	<.001***	-1.97

Note. Normative Mean (M=100), *$p < .05$. **$p < .01$. ***$p < .001$

Table 11

Comparison of SLI Group to Normative Mean by C-LIM Cell

C-LIM Cell	n	m	SD	*t*	*df*	*p*	*Cohen's d*
Cell 1 - low/low	9	90.56	11.58	-2.45	8	.040*	-.82
Cell 2 - low/mod	13	86.91	10.66	-4.43	12	<.001	-1.23
Cell 3 - low/high	13	80.38	12.07	-5.86	12	<.001	-1.63
Cell 5 - mod/mod	14	85.12	13.49	-4.13	13	.001	-1.10
Cell 6 - mod/high	10	76.96	9.06	-8.04	9	<.001	-2.54
Cell 9 - high/high	12	77.15	15.85	-4.99	11	<.001	-1.44

Note. Normative Mean (M=100), *$p < .05$. **$p < .01$. ***$p < .001$

Similarly, when looking at the results by classified disability group when compared to the normative mean, the results by Cell mirror the results by Level in Tables

7 and 8, except Cells 3 and 5 that comprise Level 3. The results of the sample deemed eligible for a Specific Learning Disability were significant at the .001 alpha level for all C-LIM Cells, with effect sizes ranging from moderate (-.58) to large (-2.15). The mean value of SLD Cell 3 scores (M= 83.49, SD=10.20) was significantly lower than the normative mean performance on Cell 3 tests; t(73)= -13.92, p= <.001. The effect size *d* of -1.62 indicates a large effect. The mean value of SLD Cell 5 scores (M= 85.86, SD=14.34) was significantly lower than the normative mean performance on Cell 3 tests; t(77)= -8.71, p= <.001. The effect size *d* of -.99 indicates a large effect.

The test performance for students deemed eligible for classification of a Speech Language Impairment were all significantly different from the normative mean for all C-LIM Cells, with large effect sizes ranging from -.82 to -2.54. The mean value of SLI Cell 3 scores (M= 80.38, SD=12.07) was significantly lower than the normative mean performance on Cell 3 tests; t(12)= -5.86, p= <.001. The effect size *d* of -1.63 indicates a large effect. The mean value of SLI Cell 5 scores (M= 85.12, SD=13.49) was significantly lower than the normative mean performance on Cell 5 tests; t(13)= -4.13, p= .001. The effect size *d* of -1.10 indicates a large effect. Complete results for SLD and SLI disability groups compared to normative mean by C-LIM cell are in Table 10 and 11 respectively.

As expected, English Language Learners referred for a suspected learning disability or speech language impairment performed significantly lower than the normative sample mean. Means of each level decrease, as the cultural loading and linguistic demand increase, marked by the increases in defined C-LIM Levels. Notably, the effect size increases as the degree of cultural and linguistic demand increases in each level, with the highest effect size at Level 4 (*d*=2.54) which has high language demand and moderate cultural loading. The effect size on Level 4 is higher than Level 5 across all sample groups in which they differ on degree of cultural loading, and size of variability in the sample. Level 4 (SD=9.06) has a consistently larger standard deviation than Level 5 (15.85) across groups, especially for the Speech Language Impairment group where there is nearly a 4 point difference in standard deviation. The increase in effect size with increasing level demonstrates the impact of cultural and linguistic factors on test performance as a function of what the tests are measuring (speech, language, cognition) and how much language proficiency is necessary for normative performance. As a result, performance on tests used in speech language evaluations are greatly affected by the degree to which person differs in their English language development as compared to monolingual English speakers.

Evaluation of Hypothesis 2

To examine the difference between the sample of ELLs referred for evaluation, and the predicted pattern of performance of typically developing ELLs based on empirical research, additional one-sample t-tests were conducted. The procedures of analysis were repeated by comparison of mean of each C-LIM level to the expected mean for an ELL with a moderate difference in cultural and linguistic experiences. Whereas the first analysis compared each level and all cells to the normative mean of 100, in this analysis each one-sample t-test for the five levels contained a different comparison value. The comparison values by level as determined in Dynda (2008) for a moderately different English Language Learner were Level 1= 94, Level 2=91, Level 3 = 87, Level 4 = 82, and Level 5=77.

Results indicated that comparison of the five C-LIM Levels of cultural and language loading to the expected performance for English Language Learners were statistically significantly for Levels 2,3,4, and 5, with p values ranging from $<.001$ to .024, and effect sizes ranging from small (-.25) to moderate (-.67). The results of the overall sample are listed by Level and Cell in Tables 12 and 13, by the SLD group Levels and Cells in Tables 14 and 15, and by SLI group on Tables 16 and 17. The mean value of Level 1 scores ($M=92.44$, $SD=12.27$) was not significantly lower than the comparison value of performance (94) for English Language Learners on Level 1 tests; $t(40)= -.82$, $p=.420$. Since the p-value for this test is larger than .05, there is no significant difference between the two values compared, and as such the effect size, Cohen's d is not interpretable. The mean value of Level 2 scores ($M=83.50$, $SD=11.17$) was not significantly lower than the comparison value of performance (91) for English Language Learners on Level 2 tests; $t(75)= -5.58$, $p<.001$. The effect size *d* of -.67 indicates a moderate effect. The mean value of Level 3 scores ($M=84.22$, $SD=8.54$) was not significantly lower than the comparison value of performance (87) for English Language Learners on Level 3 tests; $t(89)= -3.09$, $p=.003$. The effect size *d* of -.33 indicates a small effect. The mean value of Level 4 scores ($M=79.69$, $SD=9.10$) was not significantly lower than the comparison value of performance (82) for English Language Learners on Level 4 tests; $t(81)= -2.30$, $p=.024$. The effect size *d* of -.25 indicates a small effect. The mean value of Level 5 scores ($M=79.80$, $SD=10.07$) was significantly lower than the comparison value of performance (77) for English Language Learners on Level 5 tests; $t(77)=2.46$, $p=.016$. The effect size *d* of .28 indicates a small effect. Complete results comparing C-LIM Level scores to comparison values expected of moderately different bilingual students are in Table 12.

Table 12

Comparison of Overall Sample to Moderately Different ELL Scores by C-LIM Level

C-LIM Levels	M-Ex	n	M-Ob	SD	*t*	*df*	*p*	*Cohen's d*
Level 1 - low/low	94	41	92.44	12.27	-.82	40	.420	-.13
Level 2 - low/mod	91	76	83.50	11.17	-5.58	75	<.001***	-.67
Level 3 - low/high	87	90	84.22	8.54	-3.09	89	.003**	-.33
Level 4 - mod/high	82	82	79.69	9.10	-2.30	81	.024*	-.25
Level 5 - high/high	77	78	79.80	10.07	2.46	77	.016*	.28

Note. *$p < .05$. **$p < .01$. ***$p < .001$, Cohen's *d* uninterpretable if, p > .05.

Table 13

Comparison of Overall Sample to Moderately Different ELL Scores by C-LIM Cell

C-LIM Cell	M-Ex	n	M-Ob	SD	*t*	*df*	*p*	*Cohen's d*
Cell 1 - low/low	94	41	92.44	12.27	-.82	40	.420	-.13
Cell 2 - low/mod	91	76	83.50	11.17	-5.85	75	<.001***	-.67
Cell 3 - low/high	87	82	83.15	10.51	-3.32	81	.001***	-.37
Cell 5 - mod/mod	87	86	86.24	14.60	-.49	81	.63	-.05
Cell 6 - mod/high	82	82	79.69	9.10	-2.30	85	.024 **	-.25
Cell 9 - high/high	77	78	79.80	10.07	2.46	77	.016 **	.28

Note. *$p < .05$. **$p < .01$. ***$p < .001$, Cohen's *d* uninterpretable if, p > .05.

Results evaluating the difference between obtained scores of the overall sample by C-LIM cells and the evidence based comparison value of expected performance for a moderately different English language learner indicate significant differences for four of six cells with available data and effect sizes ranging from negligible (-.05) to moderate (-.67). Because there is only data in Cells 1,2,3,5, and 6, results of Cell 1 are equivalent to Level 1 as usual, Cell 2 is equivalent to Level 2 in the absence of Cell 4, Cell 6 is equivalent to Level 6 in the absence of cell 8, and Cell 9 is equivalent to Level 5 as usual.

Therefore the only C-LIM Level with further C-LIM cell analysis is Level 3 with Cells 3 and 5. The mean value of overall Cell 3 scores (M= 83.15, SD=10.51) was significantly lower than the comparison value (87) on Cell 3 tests; t(81)= -3.32, p= <.001. The effect size *d* of -.37 indicates a small effect. The mean value of overall Cell 5 scores (M= 86.24, SD=14.60) was not significantly lower than the comparison value of (87) on Cell 5 tests; t(81)= -.49, p=.63. As there is no significant difference between the obtained and expected scores, the effect size is not interpretable. Complete results of C-LIM Cell scores compared to expected of moderately different bilingual students are in Table 13.

In a comparison of bilingual student performance eligible for classification of SLD, to the expected scores of moderately different ELLs, four of five C-LIM Level t-tests indicate a statistically significant difference, with effect sizes ranging from small to moderate (-.09 to -.65). The mean value of Level 1 scores (M=92.84, SD=12.41) for students classified with SLD was not significantly lower than the comparison value of expected performance (94) for moderately different ELLs on Level 1 tests; t(36)= -.57, p= .573. As there is no significant difference between the obtained and expected scores, (p>.05) the effect size is not interpretable. The mean value of Level 2 scores (M=83.54, SD=11.51) for students classified with SLD was significantly lower than the comparison value of expected performance (91) for moderately different ELLs on Level 2 tests; t(68)= -5.39, p<.001. The effect size *d* of -.65 indicates a moderate effect. The mean value of Level 3 scores (M=82.25, SD= 8.77) for students classified with SLD was significantly lower than the comparison value of expected performance (87) for moderately different ELLs on Level 3 tests; t(81)= -2.84, p=.006. The effect size *d* of -.31 indicates a small effect. The mean value of Level 4 scores (M=79.85, SD= 9.36) for students classified with SLD was significantly lower than the comparison value of expected performance (82) for moderately different ELLs on Level 4 tests; t(75)= -2.00, p=.049. The effect size *d* of -.23 indicates a small effect. The mean value of Level 5 scores (M=79.64, SD= 10.31) for students classified with SLD was significantly different than the comparison value of expected performance (77) for moderately different ELLs on Level 5 tests; t(70)= 2.16, p=.034. The effect size *d of -.26* indicates a small effect. The difference was unexpectedly higher than the expected performance for an ELL student with a moderate difference in their language proficiency and cultural exposure. Complete results are in Table 14.

Table 14

Comparison of SLD Group to Moderately Different ELL Scores by C-LIM Level

C-LIM Levels	M-Ex	n	M-Ob	SD	*t*	*df*	*p*	*Cohen's d*
Level 1 - low/low	94	37	92.84	12.42	-.57	36	.57	-.09
Level 2 - low/mod	91	69	83.54	11.51	-5.39	68	<.001***	-.65
Level 3 - low/high	87	82	84.25	8.77	-2.84	81	.006**	-.31
Level 4 - mod/high	82	76	79.85	9.36	-2.00	75	.049*	-.23
Level 5 - high/high	77	71	79.64	10.31	2.16	70	.034*	.26

Note. *$p < .05$. **$p < .01$. ***$p < .001$, Cohen's *d* uninterpretable if, p > .05.

Table 15

Comparison of SLD Group to Moderately Different ELL Scores by C-LIM Cell

C-LIM Levels	M-Ex	n	M-Ob	SD	*t*	*df*	*p*	*Cohen's d*
Cell 1 - low/low	94	37	92.84	12.42	-.57	36	.57	-.09
Cell 2 - low/mod	91	69	83.54	11.51	-5.39	68	<.001***	-.65
Cell 3 - low/high	87	74	83.49	10.20	-2.96	73	.004**	-.34
Cell 5 - mod/mod	87	78	85.86	14.34	-.70	77	.49	-.08
Cell 6 - mod/high	82	76	79.85	9.36	-2.00	75	.049*	-.23
Cell 9 - high/high	77	71	79.64	10.32	2.16	70	.034*	.26

Note. *$p < .05$. **$p < .01$. ***$p < .001$, Cohen's *d* uninterpretable if, p > .05.

Results evaluating the difference between obtained scores of the SLD group by C-LIM cells and the evidence-based comparison value of expected performance for a moderately different English Language Learner indicate significant differences for four of six cells with available data. Cells 1, 2, 6, 9 are equivalent to their corresponding C-LIM Levels as there is only one cell of data available for each. Therefore the only C-LIM Level with further C-LIM cell analysis is Level 3 which has data in Cells 3 and 5. The mean value of SLD Cell 3 scores (M= 83.49, SD=10.20) was significantly lower than the comparison value (87) on Cell 3 tests; t(73)= -2.96, p= .004. The effect size *d* of -.34 indicates a small effect. The mean value of SLD Cell 5 scores (M= 83.86, SD=14.34) was not significantly lower than the comparison value (87) on Cell 5 tests; t(77)= -.70, p= .49. As there is no significant difference between the obtained and expected scores (p>.05) the effect size is not interpretable.. Complete results of C-LIM Cell scores compared to expected performance of moderately different bilingual students are in Table 15.

A comparison of bilingual student's performance who are eligible for classification of an SLI, to the expected scores of moderately different ELLs, indicate no statistically significant differences on all C-LIM Level t-scores. As such, the Cohen’s *d* that would indicate the size of the difference, is not interpreted for all five C-LIM levels. The mean value of Level 1 scores (M=90.56, SD=11.58) for students classified with SLI was not significantly different from the comparison value of expected performance (94) for moderately different ELLs on Level 1 tests; t(8)= -.90, p= .40. The mean value of Level 2 scores (M=86.91, SD=10.66) for students classified with SLI was not significantly different from the comparison value of expected performance (91) for moderately different ELLs on Level 2 tests; t(12)= -1.38, p=.19. The mean value of Level

3 scores (M=82.41, SD=10.36) for students classified with SLI was not significantly different from the comparison value of expected performance (87) for moderately different ELLs on Level 3 tests; t(13)= -1.66, p=.12. The mean value of Level 4 scores (M=76.96, SD=9.06) for students classified with SLI was not significantly different from the comparison value of expected performance (82) for moderately different ELLs on Level 4 tests; t(9)= -1.76, p=.11. The mean value of Level 5 scores (M=77.15, SD=15.85) for students classified with SLI was not significantly different from the comparison value of expected performance (77) for moderately different ELLs on Level 5 tests; t(11)= .03, p=.97. Complete results of SLI students by C-LIM level are in Table 16.

Table 16

Comparison of SLI Group to Moderately Different ELL Scores by C-LIM Level

C-LIM Levels	M-Ex	n	M-Ob	SD	*t*	*df*	*p*	*Cohen's d*
Level 1 - low/low	94	9	90.56	11.58	-.89	8	.40	-.30
Level 2 - low/mod	91	13	86.91	10.66	-1.38	12	.19	-.38
Level 3 - low/high	87	14	82.41	10.36	-1.66	13	.12	-.44
Level 4 - mod/high	82	10	76.96	9.06	-1.76	9	.11	-.56
Level 5 - high/high	77	12	77.15	15.85	.03	11	.97	.01

Note. **p* < .05. ***p* < .01. ****p* < .001, Cohen's *d* uninterpretable if, p > .05.

Table 17

Comparison of SLI Group to Moderately Different ELL Scores by C-LIM Cell

C-LIM Levels	M-Ex	n	M-Ob	SD	*t*	*df*	*p*	*Cohen's d*
Cell 1 - low/low	94	9	90.56	11.58	-.89	8	.40	-.30
Cell 2 - low/mod	91	13	86.91	10.66	-1.38	12	.19	-.38
Cell 3 - low/high	87	13	80.38	12.07	-1.98	12	.07	-.55
Cell 5 - mod/mod	87	14	85.11	13.49	-.52	13	.61	-.14
Cell 6 - mod/high	82	10	76.96	9.06	-1.76	9	.11	-.56
Cell 9 - high/high	77	12	77.15	15.85	.03	11	.97	.01

Note. **p* < .05. ***p* < .01. ****p* < .001, Cohen's *d* uninterpretable if, p > .05.

Results evaluating the difference between obtained scores of the SLI group by C-LIM cells and the evidence-based comparison value of expected performance for a moderately different English Language Learner, t-tests indicate there are no statistically significant differences. Cells 1, 2, 6, 9 are equivalent to their corresponding C-LIM Levels as there is only one cell of data available for each. Therefore the only C-LIM Level with further C-LIM cell analysis is Level 3 which has data in Cells 3 and 5. The mean value of SLI Cell 3 scores (M= 80.38, SD=12.07) was not significantly lower than the comparison value (87) on Cell 3 tests; t(12)= -1.98, p= .07. As there is no significant difference between the obtained and expected scores, the effect size is not interpretable. The mean value of SLD Cell 5 scores (M= 85.12, SD=13.49) was not significantly lower than the comparison value (87) on Cell 5 tests; t(13)= -.52, p= .61. As there is no significant difference between the obtained and expected scores, the effect size is not interpretable. Complete results of C-LIM Cell scores compared to expected of moderately different bilingual students are in Table 17.

When comparing the performance of the overall referred sample to the expected performance of moderately different English Language Learners with otherwise average functioning, there were significant differences ranging from <.001 to .049, and smaller effect sizes than when compared to the normative sample. The most significant difference was on Level 2 which maintained the significant difference at the .001 alpha level. The scores from the referred sample, eligible for SLD follow this pattern with significant differences on Levels 2,3,4,5 with decreasing effect sizes ranging from moderate (-.65) to small (.23) as culture and language demands increase. Students eligible for SLD scored lower than the predicted score, with the exception of Level 5 in which performance was higher.

While students eligible for SLI scored consistently lower than the predicted scores for ELL students on Levels 1, 2, 3 and 4, there were no statistically significant differences between the obtain performance and the expected scores of moderately different English Language Learners. The Level 5 scores were the only level that met the predicted score expectation (M=77.15). Results indicate that those categorized with eligibility for a Speech Language Impairment were scored in a declining pattern with scores attenuated by language and culture as expected by the C-LIM for students with a moderate language difference.

Additional Analyses

The means of C-LIM Levels for this sample appear to follow a generally linear decline along the predicted scores for moderately different ELLs, except for Level 2. To better understand the significant differences emerging for Level 2, with scores that were significantly different from the normative mean and from the moderate ELL predicted scores, with much lower than expected means, further analyses were conducted. Noting significant differences in the expected pattern of performance and the observed pattern of scores on Cell 2 and Level 2, the cells were analyzed for means deviating from the expected mean. The C-LIM Level 2 is composed of Cells 2 and 4. In this sample there were tests classified only in Cell 2. The expected mean of this cell for the moderately different sample is 91, and the mean for the whole sample at Level 2 was M=82.50; SD=11.17) is significantly lower ($p < .001$). The subtests that compose Cell 2 in this sample are CTOPP-2 Memory for Digits (M=78.30), WISC-V Block Design (M=92.18), WISC-V Picture Span (M=87.32), WISC-V Coding (M=89.50), WISC-V Symbol Search (M=92.30). An inspection of the subtests that compose Cell 2 in this sample, suggested that of all the subtests in Level 2, CTOPP-2 Memory for Digits (CTOPP-MdF) was identified as an outlier with a mean 9 points lower than the next lowest scoring subtest in the Cell.

When CTOPP-MfD is removed from the aggregate Level 2 calculation, the mean increases within 2 points of the expected performance of a moderately different ELL student. With the Level 2 mean score increase in the overall sample (M=89.93; SD=8.07), the pattern of scores is aligned with the expected scores of moderately different ELLs. Such an increase in the Cell 2 and Level 2 mean, caused changes in the statistical comparisons in each of the hypotheses. The Cell two comparison to the normative mean is still statistically significant, with a lower effect size. The Cell 2 comparison to the moderate ELL predicted score is no longer significant, decreasing from $p<.001$ to $p=.396$ in the overall sample. This supports the informal hypothesis that removing CTOPP-MfD from Cell 2 would result in a pattern consisted with the decline of scores of a moderately different ELL student. Similar changes are present for the SLD and SLI groups, as t-tests are no longer significant and effect sizes are much smaller. The post-hoc changes in Cell 2 means, as well as t-tests compared to the normative mean are in Table 18, and the t-tests comparing scores to the predicted scores for moderately different ELLs are in Table 19.

The subtest, CTOPP-MfD, may be a better fit in C-LIM Level 3 in Cell 3, or C-LIM Level 4 in Cell 6. This informal hypothesis was tested by recalculating the Level 3 and Level 4 means with CTOPP-MfD assigned to Cell 3 and Cell 6. Means of the affected C-LIM cells and levels are available in Table 18 and 19, along with the post-hoc analyses for each study hypothesis, accounting for removal of CTOPP-MfD from Cell 2 and re-classification to Cell 3 or 6.

Table 18

Post-Hoc One-Sample t-Test Results for C-LIM Cells and Levels with Proposed Test Re-Classification: Normative Mean Comparison

		Cell Means							
C-LIM Levels/Cells	M-Ex	M-Ob	SD	M-PH	SD	*t*	*df*	*p*	*Cohen's d*
Level 2/Cell2	100	83.50	11.17	89.93	8.07	-8.08	41	<.001	-1.25
SLD Level 2/Cell 2	100	83.54	11.51	90.74	7.93	-7.10	36	<.001	-1.17
SLI Level 2/Cell 2	100	86.91	10.66	92.18	9.65	-2.43	8	.041*	-0.81
Level 3	100	84.22	8.53	83.31	8.55	-18.62	90	<.001	-1.95
Cell 3	100	83.15	10.51	80.74	10.30	-17.63	88	<.001	-1.87
SLD Level 3	100	84.25	8.77	83.25	8.87	-17.38	82	<.001	-1.91
SLD Cell 3	100	83.49	10.20	80.94	10.08	-17.02	80	<.001	-1.89
SLI Level 3	100	82.41	10.36	81.85	10.20	-6.66	13	<.001	-1.78
SLI Cell 3	100	80.38	12.07	78.99	11.33	-6.94	13	<.001	-1.86
Level 4/Cell 6	100	79.69	9.10	79.59	8.98	-20.58	81	<.001	-2.27
SLD Level 4/ Cell 6	100	79.85	9.36	79.75	9.21	-19.17	75	<.001	-2.20
SLI Level 4/ Cell 6	100	76.96	9.06	77.19	8.56	-8.43	9	<.001	-2.67

Note. **p* < .05. ***p* < .01. ****p* < .001. CTOPP Memory for Digits' proposed removal from Cell 2 (Level 2) and re-classification to Cell 3 or Cell 6.

Table 19

Post-Hoc One-Sample t-Test Results for C-LIM Cells and Levels with Proposed Test Re-Classification: Moderately Different ELL Comparison

		Cell Means							
C-LIM Levels/Cells	M-Ex	M-Ob	SD	M-PH	SD	*t*	*df*	*p*	*Cohen's d*
Level 2/Cell2	91	83.50	11.17	89.93	8.07	-.86	41	.40	-.13
SLD Level 2/Cell 2	91	83.54	11.51	90.74	7.93	-.20	36	.85	-.03
SLI Level 2/Cell 2	91	86.91	10.66	92.18	9.65	.37	8	.72	.12

Level 3	87	84.22	8.53	83.31	8.55	-4.12	90	<.001	-.43
Cell 3	87	83.15	10.51	80.74	10.30	-5.73	88	<.001	-.61
SLD Level 3	87	84.25	8.77	83.25	8.87	-3.89	82	<.001	-.43
SLD Cell 3	87	83.49	10.20	80.94	10.08	-5.41	80	<.001	-.60
SLI Level 3	87	82.41	10.36	81.85	10.20	-1.89	13	.08	-.51
SLI Cell 3	87	80.38	12.07	78.99	11.33	-2.65	13	.02*	-.71
Level 4/Cell 6	82	79.69	9.10	79.59	8.98	-2.43	81	.02*	-.27
SLD Level 4/ Cell 6	82	79.85	9.36	79.75	9.21	-2.13	75	.04*	-.24
SLI Level 4/ Cell 6	82	76.96	9.06	77.19	8.56	-1.78	9	.11	-.56

Note. $*p < .05$. $**p < .01$. $***p < .001$. CTOPP Memory for Digits' proposed removal from Cell 2 (Level 2) and re-classification to Cell 3 or Cell 6.

The CTOPP Memory for Digits subtest was producing lower values than expected in Level 2, and consideration for where it may be better classified was explored by moving it to Cell 3 and Cell 4. The C-LIM, Cell 3 is defined as low cultural loading, and high linguistic demand. In this sample Cell 3 includes similar subtests as far as test constructs measuring short-term memory, with verbal stimulus of numbers and may be a better fit to re-classify CTOPP-MfD as a Cell 3, and Level 3 subtest. In this sample the other Cell 3 subtests are TAPS-4 Number Memory Forward, CTOPP-2 Rapid Digit Naming, and WISC-V Digit Span. When CTOPP-MfD is moved to Cell 3 the mean of the Cell in the overall sample, in the SLD and SLI groups decreased by 2-3 points. Again, the statistical analyses were repeated with the subtest in Cell 3, and results indicate comparisons to the normative mean are still significant (p<.001) with a large effect size (*d*=1.869) in the overall sample and similar results for each of the SLD and SLI groups. With CTOPP-MfD in Cell 3, comparisons to the moderately different ELL score in Level 3 and Cell 3 for the overall sample and SLD group remain significant (p<.001) with an increase in effect size in both groups. The SLI group comparison changed from insignificant to significant (p=.020) and increase in effect size from -.549 to -.707 in Cell

3. The post-hoc t-tests compared to the normative mean are in Table 18, and the t-tests comparing scores to the predicted scores for moderately different ELLs are in Table 19. Figure 1 illustrates the change in pattern of scores with the recommended re-classification for CTOPP-MfD to Cell 6 in the post -hoc analysis and impact to overall scores.

The other cell CTOPP-2-Memory for Digits (M=78.30) may fit is Cell 6 in Level 4 with a comparable mean of 79.70. The other tests in this cell are most of the CTOPP-2 subtests that measures phonological processing, and have a moderate cultural loading and high language demand. When CTOPP-MfD is moved to Cell 6 the Cell mean doesn't change significantly as it decreases to 79.58. The comparison to the normative does not yield changes to significance as it remains $p<.001$ and very small changes to the effect size from -2.231 to -2.273. The comparison to the moderately different ELL score the overall scores and the SLD group is still significantly different ($p<.05$) and effect sizes increased very little by .015. The comparison for the SLI group is remains insignificant and the effect size increased by .006. The post-hoc t-tests compared to the normative mean are in Table 18, and the t-tests comparing scores to the predicted scores for moderately different ELLs are in Table 19.

Additional analyses were conducted comparing mean scores obtained for each C-LIM level to each of the other levels to establish whether significant differences exist between the levels as they are designed to be differentiated by varying degrees of cultural loading and linguistic demand. These comparisons were conducted including the post-hoc re-classification of the CTOPP Memory for Digits subtest out of Level 2(post-hoc) and into Level 3(post-hoc). Results of paired t-test comparisons between C-LIM indicated that there are significant differences between the Levels for all comparisons except for Level 1 with Level 2, and Level 4 with Level 5. In this sample Level 4 only includes cell 6 with moderate culture loading and high language demands which is not very different from Cell 9 high culture/high language. This could be a contributing factor to the lack of significant difference between adjacent cells of Level 4 and Level 5.

Table 20

Paired Samples t-Tests Comparisons of the C-LIM Levels for the Overall Sample

C-LIM Level Pairs	M-Diff	SD	*t*	*df*	*p*	*Cohen's d*
Level 1 – Level 2(PH)	2.21	13.00	1.01	40	.283	.17
Level 1 – Level 3	6.47	12.08	3.43	40	<.001***	.54
Level 1 – Level 4	9.14	17.04	2.98	30	.003**	.54
Level 1 – Level 5	9.67	11.22	5.25	36	<.001***	.86
Level 2(PH) – Level 3(PH)	4.56	9.60	3.08	41	.004**	.48
Level 2(PH) – Level 4	7.25	10.33	3.97	31	<.001***	.70
Level 2(PH) – Level 5	8.22	11.26	4.50	37	<.001***	.73
Level 3(PH) – Level 4	3.38	8.32	3.65	80	<.001***	.40
Level 3(PH) – Level 5	3.98	9.97	3.48	75	<.001***	.40
Level 4 – Level 5	1.61	10.80	1.24	68	.110	.15

Note. $^*p < .05$. $^{**}p < .01$. $^{***}p < .001$. CTOPP Memory for Digits' proposed removal from Cell 2 (Level 2) and re-classification to Cell 3(Level 3).

DISCUSSION

Evaluation of Hypotheses

The first hypothesis predicted that the performance of English Language Learners referred for Speech Language Impairment would perform significantly lower on tests measuring cognitive abilities, and speech-language abilities, in comparison to the normative mean (M=100), as a proxy for comparison to monolingual English-speaking students. Scores were examined in two stages, first the overall sample of referred students, then by disability group students were determined to be eligible for (e.g. Specific Learning Disability or Speech Language Impairment). Results supported the first hypothesis because all comparisons to the normative mean across all groups, and all C-LIM levels were significant.

Results from the comparison to the normative mean were that the overall referred sample, mean scores for are significantly lower than the normative means at p= <.001 for all C-LIM levels. The effect sizes for the overall sample group increase steadily from Level 1 (d=-.62) to Level 4 (d = -2.2) and the effect size decreases a bit on Level 5 (d = -2.01) as combined cultural and linguistic demands increase. The results in the SLD group comparison to the normative mean were similarly significant across all C-LIM levels at the p=.001 or <.001. The effect sizes for the SLD group demonstrate and increasing pattern from Level 1 (d = -.57) to Level 4 (-2.15) and again decreasing to (-1.98) as the overall sample did.

Notable findings from the comparisons to the normative mean in the SLI group supported the hypothesis that the performance for students suspected of a speech-language impairment would score significantly lower than the normative mean across all

C-LIM Levels with $p < .05$ on Level 1 and $p < .001$ on the remaining Levels. The effect size increased as the cultural and language demand increased from C-LIM Level 1 (-.82) to Level 4 (-2.54) and decreased on Level 5 (-1.442). There is an evident decline in the performance of SLI ELLs relative to the subtest students are administered, based on C-LIM classification of subtests as low/high language and low/high culture demands.

Though the means on Levels 4 (76.96) and 5 (77.15) are nearly the same, the effect size is notably lower on Level 5. A possible explanation is the difference in variability of scores on the two levels as Level 4 has an SD=9.06 while Level 5 has a SD=15.85 which is a third larger. A closer look at the range of Level 5 mean scores is 45 points, with a minimum of 55 and maximum scores of 100, while the range of Level 4 is 24.17 points, with a minimum of 67.50 and a maximum score of 91.67. The lowest score in Level 5 was lower than any score in Level 4 and the highest score in Level 5 is also higher than any Level 4 score. This difference in the range of the sample in each Level explains why the effect size is lower for SLI Level 4 despite a similar mean to Level 5. The mean scores obtained for Level 4 and Level 5 are just .19 points different and are not statistically different ($p=.11$). It is possible students in the present sample have a range of English Proficiency that is below proficient to average, but not so low in proficiency that the scores would be much lower for tests with high language loading. Without further background, it is also possible the sample doesn't contain very low English proficient students due to deferred testing until students receive more English instruction. In fact, the student sample scored just as expected of a moderately different English Language Learners in Level 5 (M=77).

As expected, English Language Learners performed significantly lower than the normative mean for standardized testing used for cognitive and speech-language assessments. The values of the sample means are relatively consistent with the expected performance of EL students, with the exception of Level 2, and is strong support that performance is affected by language, in a manner that is consistent with C-LIM classifications.

The second hypothesis predicted that there would not be a significant difference between the performance of English Language Learners with an identified Speech Language Impairment, and the expected performance for students considered moderately different in their acculturation and language development. A non-significant difference supports the hypothesis that English Language Learners referred for an SLI are performing as predicted by prior research and scores are attenuated as a function of their degree of language experience including their cultural experiences and formal education in English.

Results for the overall sample and the SLD group comparisons had significant differences ($p<.05$) between performance on cognitive and speech testing relative to the predicted scored for moderately different English Language Learners for C-LIM Levels 2 through 5. The most significant difference in the overall and SLD samples was on Level 2 ($p<.001$) with a moderate effect size (-.671; -.649). For the SLD group, mean scores on the C-LIM Levels begin as expected of a moderately different ELL with Level 1 (low/low) having no significant difference from the predicted score. The score means for Level 1, 2, 3, and 4, scores are lower than the predicted scores, in a generally declining pattern and on Level 5 are higher than predicted. Though the pattern of scores follow a general declining pattern suggestive of a language difference rather than a disability, these aggregated results of 103 students eligible for SLD may be reflective of the variability of specific learning disabilities represented in the sample, with strengths and weaknesses in different domains.

Results for the SLI group comparisons were not significantly different from the predicted scores of ELLs with a moderate language difference on cognitive and speech testing used in SLI assessments (WISC-V, CTOPP-2, TAPS-4). Scores for the 20 students determined to be eligible for SLI, were lower than the predicted scored for moderately different ELLs, but the differences did not reach statistical significance. The C-LIM Level with highest cultural loading and highest language demand, Level 5 demonstrated a score (M=77.15) as expected from moderately different ELLs at this level (M=77). It is possible the high variability in this group for Level 5 scores (SD=15.85) is due to a range of reasons for speech-language referrals that may be other speech focused weaknesses not measured by the tests analyzed such as articulation. The pattern of scores across the C-LIM levels is not statistically significantly different from the expected scores of ELLs without disabilities, and considered to have a moderate language difference. The hypothesis was supported and the results have important implications for evaluation of English Language Learners referred for Speech Language Impairments considered to have moderately different acculturative experience and language instruction from their monolingual mainstream culture peers.

Study Limitations and Future Directions

The study findings add to the growing research on the utility of the Culture-Language Interpretive Matrix, and the need to carefully consider cultural and language differences when testing English Language Learners to determine educational disabilities or mere language differences. Study provides a preliminary look at patterns of performance for English Language Learners tested for suspected Speech Language Impairment who have had significantly different cultural and linguistic experiences relative to their monolingual peers. There are limitations of the study that offer future directions for empirical research in this area.

A primary limitation is that this study utilized archival records from previously completed psychoeducational evaluations in Central California. The existing student cases collected from this area were identified as entirely Hispanic, with Spanish listed as either primary or second language. This feature of the sample may limit generalizability to other diverse groups of students of other cultures and languages spoken. The archival nature of the data collection, via another ongoing study limited sampling for cases with students referred for Speech Language Impairments and resulted in a small sample size for the group targeted for examination. Due to the prevalence of evaluations school psychologists are involved with for students referred for Speech Language Impairment, it would be beneficial for future researchers to collect larger sample sizes, via a state-wide sample. A larger sample may also be representative of various factors that impact degree of language difference including parental education, time in the U.S and exposure to acculturative experiences. With the small sample size there was great variability in the tests used for speech-language testing and the sample for each test used was smaller still. Some language tests broadly used in evaluations (e.g., CASL, CELF, Goldman-Fristoe) were not included as they did not meet the minimum frequency for inclusion in the study.

A larger and controlled study that may increase consistency of tests used for evaluation would yield a higher sample for increased power in analysis.

As testing data were collected from evaluations conducted by various evaluators, the factors of test selection, administration, and interpretation where not controlled and therefore inconsistent. Some test batteries were not consistently administered in full, such as the WISC-V, some students were given the core 7 subtests, while others were administered the complete 10 subtests standard battery. This decreased the amount of data available in each C-LIM cell and contributed to differences in variability. Additionally, in an effort to conduct evaluations with bilingual students, evaluators may have given credit for an answer in Spanish, or administration may have been modified and not disclosed. Deviation from standardized administration was not controlled in this study as data was collected after evaluations were completed. In future controlled studies evaluators may be trained for reliability of standardized test administration practices.

This sample is inadvertently restricted to English Learners with limited English proficiency, suggesting they would be considered moderately or markedly different. This study did not reliably measure cultural and linguistic developmental factors that determine degree of difference. There is limited information in the evaluation reports indicating developmental factors such as time in the U.S, parental backgrounds, English learning settings (i.e., pull-out English as a Second Language, Dual instruction), or family income. The data available measuring English Language Proficiency is the ELPAC assessment which was reportedly administered to 57 students. With the available ELPAC scores, made determination the sample was moderately different, but it is possible they were more markedly different. Additionally, Bilingual students who previously passed the ELPAC or were not classified as an English Learner despite being bilingual, are not in this sample. A future controlled study may require all students to be administered the same proficiency measure, and use a standard questionnaire to collect developmental data for consistent determination of degree of language difference.

This study does not have control or comparison groups as all archival data available were for English Language Learners referred for evaluation. This limitation in the participants does not allow for comparison of referred ELLs to referred monolingual English speakers, or to compare the referred English Speaker's performance on Speech Language testing to the normative mean. Future directions in research with additional groups such as a non-referred English Language Learner group, a non-referred monolingual English speaker group, and a referred monolingual English speaker group would provide interesting and important comparisons related to cultural and linguistic impacts on testing.

IMPLICATIONS FOR PROFESSIONAL PRACTICE OF SCHOOL PSYCHOLOGY

Though there are some limitations to this current study, nevertheless, there are important implications for practicing school psychologists conducting evaluations with English Language Learners in order to reduce misidentification and overrepresentation of ELLs in Special Education. Findings from this study support the body of research on the systematic impact of culture and language on ELL test performance on testing for students referred for speech-language evaluations and the utility of the C-LIM in score interpretation. This study advances knowledge and evidence of cultural and linguistic impacts on testing for those referred for language difficulties. Additionally, based on the results of this study, considerations for changes to classification of CTOPP-Memory for Digits in the C-LTC to further decrease bias and discriminatory practices in psychological testing.

A significant implication of this study is the recommended reclassification of CTOPP Memory for Digits. The tests in the extreme ends of the Culture-Language Interpretive Matrix (Low Language/Low Culture and High Language/High Culture) are easier to classify, however those in the moderate levels, are more difficult to determine. This results in some classification errors that are likely to involve tests that contain moderate cultural loading, or moderate language demand. The results of the additional analyses of this study suggest that the CTOPP Memory for digits subtest, originally categorized as a Cell 2 test, fits better with the mean performance of Cell 6 in Level 4. This change in the Cultural-Language Test Classifications, would improve the accuracy of the measured impact of culture and language, and improve interpretations of score profiles. The CTOPP-MfD appears to fit better with the test constructs at face validity with the other verbally presented tests of short-term memory in Cell 3, and developers of the C-LTC, Ortiz & Flanagan, may also consider re-classification to Cell 3 with high language and a better fit with the tests found in this sample. Developers may consider research with non-disabled ELLs to support the recommendation for reclassification.

The study sought to examine a sample of students referred for speech-language testing, to evaluate their performance against predicted performance for non-disabled English Language Learners with moderate language development differences. There weren't significant differences from the predicted trend, suggesting scores followed the expected performance of a moderately different ELL. Scores showed that the referred sample scored below the predicted scores for students considered moderately different based on language and cultural factors, though not a statistically significant difference, the scores almost matched the predicted scores for those with marked differences instead. It is possible that a referred sample with moderate difference in language is impacted to the degree they appear markedly different. This area needs additional research to determine patterns of performance of the SLI population who are ELLs.

A primary implication for practice is the recommended use of the Cultural-Language Interpretive Matrix in the interpretation of test scores in the evaluation of English Language Learners. A result of this study supports the systematic impact of culture and language, evident in the increasing effect sizes as these factors increase. This demonstrates the impact of cultural/linguistic factors on test performance as a function of what tests are measuring and how much language development/proficiency is necessary for normative performance. The C-LIM aids in mitigating discriminatory testing procedures, by reducing bias inherent of standardized tests with cultural loading and linguistic demands, that are biased toward the monolingual norming group. By using the C-LIM to assess the validity of testing results, practitioners will be able to better to distinguish between a language difference and a disability.

A subsequent implication for practice of evaluations with ELLs, when using the C-LIM is to employ intentional assessment of cultural and linguistic factors and the degree to which these will impact student scores. To accurately interpret the pattern of scores in the C-LIM, it is necessary to begin by correctly categorizing students with either slight, moderate or marked degree of language differences. To achieve this, practitioners review and collect relevant background history information, including acculturation stages, conduct a comprehensive assessment of language proficiency and development, opportunities for learning, and evaluate relevant cultural factors as part of a framework of non-discriminatory assessment (Flanagan et al., 2013). With the post-hoc correction the overall referred sample scores fall between the markedly and moderately different trend lines for Levels 1, 2, 3 and 4. C-LIM developers may consider additional degrees of language difference to more narrowly capture the differences in acculturation and language development among English Language Learners. The data in this sample show there may be a category between moderate and markedly, and in a sample with a broader range of English proficiency represented, there may be a group between moderately and slightly different. These additional categories of language difference yield five total categories of degree of language difference.

The tests examined in this study are those used in the evaluation of students referred for Speech-Language Impairments, and the use of these tests also presents issues for other practitioners administering and interpreting them, specifically Speech Language Pathologists (SLPs). It is necessary that SLPs are trained in the cultural and linguistic impact on score performance and the use of the C-LIM as a tool to evaluate the validity of test scores obtained when testing English Language Learners. By extension, trainers of SLPs should prioritize the development of a second language, and non-discriminatory assessment practices within in SLP curriculum, or continuing education training requirements. Though some SLPs train to work with Bilingual Populations, all practitioners would benefit from knowledge related to fair diagnostic assessment practice when determining if a student requires SLP services.

With the rapidly diversification of the country, practitioners need to increase training in the fair evaluation of English Language Learners, to avoid misclassification, overuse of resources (e.g., testing time) and incorrect or ineffective placement and services. Practitioners lack training in non-discriminatory testing (Ochoa et al., 2004), including comprehensive assessment of acculturative and linguistic factors. Only two states in the U.S, New York and Illinois, offer a formal Bilingual School Psychologist designation with targeted training on working with ELLs. However, all practitioners in the country are held to the APA Standards for Psychological and Educational Testing, and must seek additional trainings, professional development, and resources to adequately meet the needs of all students.

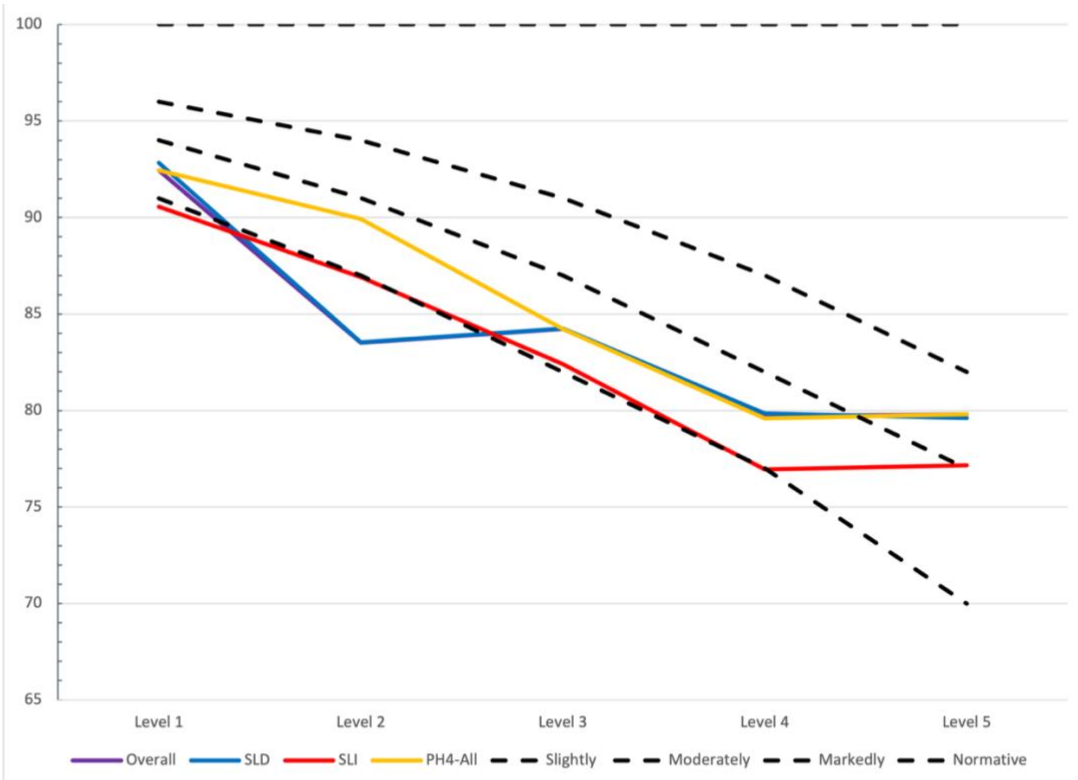
SUMMARY

The intent of the study was to determine patterns of performance of ELLS referred for Speech Language Impairment compared to predicted patterns of English Language Learner performance on cognitive and speech testing. It was anticipated that speech-language testing would have high linguistic demand resulting in low performance on the subtests, and significantly different from the normative mean.

Prior research has identified expected performance for English Learners by degree of difference. Test scores for English language learners without identified disabilities are systematically impacted by cultural loading and linguistic demands of the test, such that their scores follow a pattern of decline on the C-LIM with highest scores in Level 1 (low/low) and lowest scores in Level 5 (high/high). The extent to which a student is different from normative mainstream expectations is relative to differences in acculturative experiences and linguistic development. Researchers have identified three degrees of difference; slightly, moderate, and markedly different (Dynda, 2008; Flanagan & Ortiz, 2001; Sotelo-Dynega et al., 2013). Diverse students that are considered to be slightly different, have very high English language proficiency, lived in the U.S for over seven years, parents with high school education or above, and speak and read with native-like proficiency. Individuals considered moderately different have moderate to high English language proficiency, have lived in the U.S for 3-7 years, are able to communicate, while their parents are limited English speakers with some formal schooling, and have below grade level literacy skills. Individuals considered markedly different have low English language proficiency or limited acculturative experiences, as they may have recently arrived to the U.S within 3 years, have little to no formal education, and are beginning to develop conversational and literacy skills (Flanagan et al., 2013). The expected performance on testing with English Language Learners is attenuated by cultural and linguistic factors in an increasing fashion as C-LIM Levels increase. For example, students determined to be moderately different are expected to score less than the normative mean by 3-5 points on Level 1, 5-7 points on Level 2, 7-10 points on Level 3, 10-15 points on Level 4, and 15-20 points less on Level 5. These expected decreases occur as a function of increasing cultural and linguistic demands of tests, and if such a pattern exists, it is concluded testing is not valid to identify a disability. The C-LIM was designed as a tool to examine validity of testing, and not as a diagnostic tool. In order to more accurately determine validity and make meaningful interpretations in non-discriminatory assessment, it is necessary to adequately identify a student's degree of difference in language development.

Figure 1

Obtained Group Means, and Overall Mean with CTOPP-MfD Re-Classification to Level 4



REFERENCES

Aguera Verderosa, F. (2007). *Examining the effects of language and culture on the Differential Ability Scales with bilingual preschoolers* [Psy.D., St. John's University (New York)].

Albrecht, S. F.-G. (2011). Federal policy on disproportionality in special education: Is it moving us forward. Journal of Disability Policy Studies.

Alfonso, V. C., Bracken, B. A., & Nagle, R. J. (2020). *Psychoeducational Assessment of Preschool Children*. Routledge.

Artiles, A. R. (2005). Within-group diversity in minority disproportionate representation; English language learners in urban school districts. Exceptional Children , 71, 283- 300.

Artiles, A. J. & Klingner, J. K. (2006). Forging a knowledge base on English language learners with special needs: Theoretical, population, and technical issues. *Teachers CollegeRecord, 108,2187-2194.*

Artiles, A. J., Kozleski, E. B., Trent, S. C., Osher, D., & Ortiz, A. (2010). Justifying and Explaining Disproportionality 1968—2008: A Critique of Underlying Views of Culture. *Exceptional Children*, 22.

Barrett, C., & Matthews, M. S. (2021). Ortiz Picture Vocabulary Test: Ortiz PVAT. *The Twenty-First Mental Measurements Yearbook*.

Betz, S. K., Eickhoff, J. R., Sullivan, S. F., Nippold, M., & Schneider, P. (2013). Factors Influencing the Selection of Standardized Tests for the Diagnosis of Specific Language Impairment. *Language, Speech & Hearing Services in Schools*, *44*(2), 133–146. https://doi.org/10.1044/0161-1461(2012/12-0093)

Brown, J. E. (2008). *The use and interpretation of the Batería III with U.S. bilinguals* [Ed.D., Portland State University].

Brownell, R. (2011). Expressive One-Word Picture Vocabulary Test–Fourth Edition. Academic Therapy.

Caesar, L. G., & Kohler, P. D. (2009). Tools Clinicians Use: A Survey of Language Assessment Procedures Used by School-Based Speech-Language Pathologists. *Communication Disorders Quarterly*, *30*(4), 226–236. https://doi.org/10.1177/1525740108326334

Calderón-Tena, C. O., Styck, K. M., Vega, D., & Kranzler, J. H. (2020). Evaluating cultural and linguistic load of IQ scores for English language learners. *International Journal of School & Educational Psychology*, *0*(0), 1–17. https://doi.org/10.1080/21683603.2020.1803164

Canivez, G. L., & Graham Laughlin, T. (2021). Peabody Picture Vocabulary Test, Fifth Edition. *PPVT-5*. Mental Measurements Yearbook.

Cohen, J. (1977). The Effect Size Index. In *Statistical Power Analysis for the Behavioral Sciences* (pp. 26–27). essay, Academic Press.

Cummins, J. C. (1984). Bilingual and special education: Issues in assessment and pedagogy. Austin, TX: PRO-ED

De Valenzuela, J. C. (2006). Examining educational equity: revisiting the disproportionate representation of minority students in special education. Exceptional Children, 72 (4), 425-441.

Dhaniram-Beharry, E. (2008). *Cultural and linguistic influences on test performance: Evaluation of alternate variables* [Psy.D., St. John's University (New York)].

Diana v. St. Board of Education, C-70-37 R.F.P. (N.D. Cal. Jan. 7, 1970 and June 18,1973). Education for All Handicapped Children Act, Pub. L. No. 94-142 (1975).

Dollaghan, C. (2007). *The handbook for evidence-based practice in communication disorders*. Baltimore, MD: Paul H. Brooks Publishing Co.

Dollaghana, C. A., Horner, E. A., & Oetting, J. (2011). Bilingual language assessment: A meta-analysis of diagnostic accuracy. *Journal of Speech, Language & Hearing Research*, *54*(4), 1077-1088.

Dunn, L. M. (1968). Special education fot the mildly retarded—is much of it justifiable? *Exceptional Children, 23,5-21.*

Dunn, D. M. (2019). *Peabody Picture Vocabulary Test* (5th ed.) [Measurement instrument]. Bloomington, MN: NCS Pearson.

Dynda, A. M. (2008). The relation between language proficiency and IQ test performance. Unpublished manuscript. St. John's University, NY.

Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). The Wechsler Intelligence Scales and Gf-Gc theory: A contemporary approach to interpretation. Boston, MA: Allyn & Bacon.

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

Flanagan, D. P., Ortiz, S.O., & Alfonso, V.C. (2007). Essentials of cross-battery assessment with CD-Rom. Wiley: New Jersey. New York: Wiley

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

Harris, B., Vega, D., Peterson, L. S., & Newell, K. W. (2021). Critical Issues in the Training of Bilingual School Psychologists. *Contemporary School Psychology*, *25*(3), 273–287. https://doi.org/10.1007/s40688-020-00340-7

Harris, S. M., & Johnson, K. M. (2014). Expressive One-Word Picture Vocabulary Test, 4th Edition: EOWPVT-4. *The Nineteenth Mental Measurements Yearbook*.

Hutchins, T. L., & Pratt, S. (2017). Clinical Evaluation of Language Fundamentals–Fifth Edition: CELF-5. *The Twentieth Mental Measurements Yearbook*.

Individuals with Disability Education Act of 2004, Pub. L. No. 108-446, 118 Stat. 2647 (2004).

Irwin, V., Bailey, T. M., Panditharatna, R., Sadeghi, A., & S&P GLobal. (2024, February 28). *Elementary and Secondary Enrollment*. Projections of Education Statistics to 2030. https://nces.ed.gov/programs/PES/section-1.asp

Kilinger, J. A. (2006). English language learners who struggle with reading. Journal of Learning Disabilities, 39 (2), 108-128.

Larry P. v. Riles, 495 F. Supp. 926 (N.D. Cal. 1979).

Lella Souravlis, S. A. (2010). *Evaluating speech-language and cognitive impairment patterns via the culture-language interpretive matrix* [Psy.D., St. John's University (New York)].

Martin, N., Brownell, R., Hamaguchi, P., (2018) TAPS-4: Language Processing Skills Assessment. (4th Ed.)

McKnight, T., & Shapley, K. L. (2014). Preschool Language Scales, Fifth Edition: PLS-5. *The Nineteenth Mental Measurements Yearbook*.

McGrew, K. S., & Flanagan, D. P. (1998). The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment. Boston, MA: Allyn & Bacon

Meyer, E. L. (2013). *Diagnostic accuracy of the Culture-Language Interpretive Matrix with the WJ-III-NU: A comparison of Spanish-speaking English language learners and monolingual English-speaking students* [Ph.D., The Pennsylvania State University].

Moyle, M. J., & Newman, D. B. (2021). Comprehensive Assessment of Spoken Language, Second Edition: CASL-2. *The Twenty-First Mental Measurements Yearbook.*

National Center for Education Statistics. (2021). The condition of education 2021 (NCES 2021-508). Washington, DC: U. S. Department of Education, Institute of Education Sciences. Retrieved from https://nces.ed.gov/programs/coe/pdf/2021/cgf_508c.pdf

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* [Psy.D., St. John's University (New York)].

Ogiela, D. A., & Montzka, J. L. (2021). Norm-Referenced Language Test Selection Practices for Elementary School Children With Suspected Developmental Language Disorder. *Language, Speech & Hearing Services in Schools*, *52*(1), 288–303. https://doi.org/10.1044/2020_LSHSS-19-00067

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

Ortiz, S. O. (2001). Assessment of cognitive abilities in Hispanic children. Seminars in Speech and Language, 22(1), 17–37.

Ortiz, S. O. (2002). Best practices in nondiscriminatory assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV (pp.* 1321-1336). Washington, DC: National Association of School Psychologists.

Ortiz, S. O. (2014). Best Practices in Nondiscriminatory Assessment. In *Best Practices in School Psychology VI: Foundations* (V, Vol. 2, pp. 61–74). National Association of School Psychologists.

Ortiz, S. O. (2018). *Ortiz Picture Vocabulary Acquisition Test*. Toronto, Canada: Multi-Health Systems.

Ortiz, S. O., Piazza, N., Ochoa, S. H., & Dynda, A. M. (2018). Testing with culturally and linguistically diverse populations: New directions in fairness and validity. In *Best Practices in School Psychology VI: Foundations* (4th ed., pp. 684–712).

Rhodes, R. L., Ochoa, S. H., & Ortiz, S. O. (2005). *Assessing Culturally and Linguistically Diverse Students: A Practical Guide*. Guilford Press.

Semel, E., Wiig, E. H., & Secord, W. (2013). Clinical Evaluation of Language Fundamentals–Fifth Edition. Pearson

Skiba, R. J., Simmons, A. B., Ritter, S., Gibb, A. C., Rausch, M. K., Cuadrado, J., & Chung, C.-G. (2008). Achieving Equity in Special Education: History, Status, and Current Challenges. *Exceptional Children*, *74*(3), 264–288. https://doi.org/10.1177/001440290807400301

Sotelo-Dynega, M., Ortiz, S. O., Flanagan, D. P., & Chaplin, W. F. (2013). English Language Proficiency and Test Performance: An Evaluation of Bilingual Students with the Woodcock-Johnson Iii Tests of Cognitive Abilities. *Psychology in the Schools*, *50*(8), 781–797. https://doi.org/10.1002/pits.21706

Sotelo-Dynega, M., & Dixon, S. G. (2014). Cognitive Assessment Practices: A Survey of School Psychologists. *Psychology in the Schools*, *51*(10), 1031–1045. https://doi.org/10.1002/pits.21802

Styck, K. M., & Watkins, M. W. (2013). Diagnostic utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition among referred students. *School Psychology Review*, *42*(4), 367–382.

Sullivan, A., Kozleski, E., & Smith, A. (2008, March). Understanding the current context of minority disproportionality in special education: Federal response, state activities, and implications for technical assistance. Paper presentation at the American Educational Research Association Annual Meeting, New York, NY.

Sullivan, A. L. (2009). *Patterns and predictors of English language learner representation in special education* [Ph.D., Arizona State University].

Sullivan, A. L. (2011). Disproportionality in Special Education Identification and Placement of English Language Learners. *Exceptional Children*, *77*(3), 317–334. https://doi.org/10.1177/001440291107700304

Sullivan, A., & Bal, A. (2013). Disproportionality in Special Education: Effects of Individual and School Variables on Disability Risk. *Exceptional Children*, *79*, 475–494. https://doi.org/10.1177/001440291307900406

Tello, A. (2020). *Comparison of Language Performance Using Exposure and Non-*

exposure Norms in English Learners Identified with Speech Language Impairment [Psy.D., St. John's University (New York)].

Tychanska, J. (2009). *Evaluation of speech and language impairment using the culture - language test classifications and interpretive matrix* [Psy.D., St. John's University (New York)].

U.S. Department of Education. (2010). *The 29th annual report to Congress on the implementation of the Individuals with Disabilities Education Act, 2007, vol. 1*. Washington, DC: Office for Special Education and Rehabilitative Services, U.S. Department of Education.

Retrieved from https://nces.ed.gov/programs/PES/section-1.asp#

Vasquez, M. (2010). *Patterns of cognitive performance in preschool aged English learners with speech and language and global cognitive impairments* [Psy.D., St. John's University (New York)].

Wechsler, D., Raiford, S. E., & Holdnack, J. A. (2014). *WISC-V Technical and Interpretive Manual.* San Antonio, TX: The Psychological Corporation.

Yoo, M. S. (2015). *A validation study of the Culture-Language Interpretive Matrix and Interpretive Graph* [Psy.D., St. John's University (New York)].

Wagner, R.K.,Torgesen, J.K., Rashotte, C.A., & Pearson, N.A. (2013). CTOPP-2: Comprehensive Test of Phonological Processing. (2nd Ed). Austin,TX:PRO-ED.

Vita

Name	*Mayra Alejandra Reyes Ruiz*
Baccalaureate Degree	*Bachelor of Arts* *University of Connecticut, Storrs, CT* *Major: Psychology*
Date Graduated	*May, 2012*
Other Degrees and Certificates	*Master of Arts* *University of Connecticut, Storrs, CT* *Major: School Psychology*
Date Graduated	*August, 2016*
	Diploma in Professional Education *University of Connecticut, Storrs, CT*
Date Graduated	May, 2018