Diagnostic Utility of the Culture-Language Interpretive Matrix for the WISC-IV

Among Referred Students

by

Kara M. Styck

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2012 by the
Graduate Supervisory Committee:

Marley Watkins, Co-Chair
Roy Levy, Co-Chair
John Balles

ARIZONA STATE UNIVERSITY

August 2012

ABSTRACT

The Culture-Language Interpretive Matrix (C-LIM) is a new tool hypothesized to help practitioners accurately determine whether students who are administered an IQ test are culturally and linguistically different from the normative comparison group (i.e., different) or culturally and linguistically similar to the normative comparison group and possibly have Specific Learning Disabilities (SLD) or other neurocognitive disabilities (i.e., disordered). Diagnostic utility statistics were used to test the ability of the Wechsler Intelligence Scales for Children-Fourth Edition (WISC-IV) C-LIM to accurately identify students from a referred sample of English language learners (Ells) (n = 86) for whom Spanish was the primary language spoken at home and a sample of students from the WISC-IV normative sample (n = 2,033) as either culturally and linguistically different from the WISC-IV normative sample or culturally and linguistically similar to the WISC-IV normative sample. WISC-IV scores from three paired comparison groups were analyzed using the Receiver Operating Characteristic (ROC) curve: (a) Ells with SLD and the WISC-IV normative sample, (b) Ells without SLD and the WISC-IV normative sample, and (c) Ells with SLD and Ells without SLD. Results of the ROC yielded Area Under the Curve (AUC) values that ranged between 0.51 and 0.53 for the comparison between Ells with SLD and the WISC-IV normative sample, AUC values that ranged between 0.48 and 0.53 for the comparison between Ells without SLD and the WISC-IV normative sample, and AUC values that ranged between 0.49 and 0.55 for the comparison between Ells with SLD and Ells without SLD. These values indicate that the C-LIM has low diagnostic accuracy in terms of differentiating between a sample of Ells and the WISC-IV normative sample. Current available evidence does not support use of the C-LIM in applied practice at this time.

DEDICATION

This dissertation is dedicated to my husband and my family. Your unconditional love and support made this feat possible.

ACKNOWLEDGMENTS

I would like to thank my committee members, Dr. Watkins, Dr. Levy, and Dr. Balles, for their continued guidance and support throughout the completion of my dissertation. My dissertation work marks the beginning of my career as an applied researcher in educational psychology and I am very grateful to have had the opportunity to learn from such talented professionals.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Chapter 1

Introduction

Messick (1995) described the notion of construct validity as, "an integration of any evidence that bears on the interpretation or meaning of the test scores" (p. 742). Interpretations of test scores are considered valid when they accurately reflect the true characteristics of the examinee on the underlying construct. Construct underrepresentation and construct-irrelevant variance are two major threats to construct validity. The former occurs when test items reflect only a narrow segment of the underlying construct and the latter occurs when test items reflect other variables in addition to the underlying construct. Construct underrepresentation and construct-irrelevant variance may cause test scores to be spuriously high or low for some subgroups of test takers compared to others. Both situations are undesirable and result in invalid test score interpretations because the test items do not accurately represent the construct of interest for all examinees (Messick, 1995; Reynolds, Livingston, & Willson, 2006).

Construct validity is especially important in high stakes diagnostic decisions. The diagnosis of Specific Learning Disabilities (SLD) represents a common occurrence of this type of decision-making. SLD diagnostic decisions often rely on standardized norm-referenced IQ test score interpretations due to the widespread use of the ability-achievement discrepancy criterion (Mercer, Jordan, Allsopp, & Mercer, 1996). Students with SLD represent the majority of students served under the Individuals with Disabilities in Education Improvement Act of 2004 (IDEA; Aud et al., 2010) and approximately 2.5 million students in U.S. schools receive special education services under this educational disability category (Data Accountability Center, 2009). Weak construct validity in test score interpretations used in SLD diagnostic decisions could adversely impact millions of students.

The Impact of High-Stakes Diagnostic Decisions in the Assessment of Specific Learning Disabilities

Outcomes for students with Specific Learning Disabilities. Short- and long-term outcomes for students with SLD have been studied extensively over the past few decades through an examination of demographic characteristics of students with SLD at varying ages (e.g., Aud et al., 2010; Cortiella, 2011; Data Accountability Center, 2008-2009; Wagner et al., 2003; Wagner, Newman,

Cameto, Garza, & Levine, 2005) and through longitudinal studies that examined the ramifications of SLD diagnosis for study participants across various intervals of time (e.g., see Gottesman, 1989 for a comprehensive review of early longitudinal research; Raskind et al., 1999; Rogan & Hartman, 1990; Spekman, Goldberg, & Herman, 1992; Werner, 1993). As a group, students with SLD tend to experience poor educational (Aud et al., 2010; Data Accountability Center, 2008-2009; Wagner et al., 2003), social/emotional (Gettinger & Kosick, 2001; Higgins, Raskind, Goldberg, & Herman, 2002; Reiff, Gerber, & Ginsberg, 1997), and post-secondary outcomes (Cortiella, 2011; Wagner et al., 2005) compared to their non-identified peers.

For example, 45% of students with SLD in reading-related areas and 44% of students with SLD in math-related areas perform more than 3 grade levels below the grade in which they are enrolled (Wagner et al., 2003). Students who are diagnosed as having SLD are also less likely to graduate from high school with a regular diploma compared to their non-identified peers (Aud et al., 2010; Data Accountability Center, 2008-2009). Most disturbingly, high school drop out rates for students with SLD are more than double the high school drop out rates for non-identified students. It is estimated that 22% of students who are diagnosed as having SLD drop out of high school (Data Accountability Center, 2008-2009), whereas only 8% of students not diagnosed as having SLD engage in this behavior (Aud et al., 2010).

The act of being labeled has other implications. Higgins, Raskind, Goldberg, and Herman (2002) used qualitative research methods to identify patterns of adjustment that 41 students with SLD experienced while they reframed their disability in an adaptive manner. Participants in their study reported an awareness of the negative judgments passed on to them by society. They shared feelings of fear, confusion, frustration, and anger as a result of being labeled different from others. Students with SLD also tend to have deficits in verbal communication and social judgment, which impair their ability to make and sustain meaningful friendships (Kopp et al., 1984; Sanchez, 1984).

Post-secondary outcomes are similarly poor for students with SLD compared to their non-identified peers. Adverse outcomes for adults with SLD are most often associated with an early onset of the disability and severe impairment as a child (Bruck, 1987; Kurzweil, 1992; Schonhaut & Satz, 1983; Spreen, 1988). Recent estimates suggest that only 10% of students with SLD enroll in a four-

year college within two years of leaving high school, but 28% of students in the general education population enroll in a four-year college within the same time frame (Wagner et al., 2005). Employment post-high school or secondary school is also adversely affected by the presence of learning disabilities. Reading, writing, and/or math deficits limit career choices for adults with SLD (Finucci, 1986; Finucci et al., 1984; Sitlington & Frank, 1990) and many adults with SLD report working at a slower pace and feeling less adequate at completing job demands than their non-disabled co-workers (Haring, Lovett, & Smith, 1990). Adults with SLD are also less likely to be employed and are more likely to be unemployed partly as a result of a lack of education compared to adults who were never labeled SLD (Cortiella, 2011).

Support services for students with Specific Learning Disabilities. The accurate identification of a student as a student with SLD is important so that appropriate educational supports may be put into place. Spekman, Goldberg, and Herman (1992) followed 50 young adults with SLD who attended the Frostig Center between 1968 and 1975 for 10 years and examined protective factors related to resilience to risk for negative outcomes. They identified six characteristics of success (see Spekman et al., 1992 for detailed information) that included greater self-awareness and/or self-acceptance of the learning disability, proactivity, perseverance, emotional stability, appropriate goal setting, and the presence and use of effective social support systems. Werner (1993) identified similar protective factors for long-term success during the Kauai Longitudinal Study for a group of 22 children with SLD compared to 22 non-disabled controls.

Additionally, many academic interventions have been deemed successful at remediating learning problems. Some of these include direct instruction in specific reading skills (e.g., Begeny et al., 2010; Massetti, 2009), math skills (e.g., see Burns, Codding, Boice, & Lukito, 2010 for a review), or written language skills (e.g., see Fulk & Stormont-Spurgin, 1995; Graham & Perin, 2007 for a review) and class-wide peer tutoring programs (e.g., Delaquadri et al., 1983; Menesses & Gresham, 2009). However, some of the risk factors associated with SLD may be linked to the ineffectiveness of other interventions used with students enrolled in special education programs. Kavale and Forness (1999) synthesized the meta-analysis literature on various special education interventions. Mean effect sizes ranged from .08 to .58 for 8 classroom- and medically-based interventions including perceptual-motor

training, psycholinguistic training, modality instruction, stimulant drugs, diet intervention, psychotropic drugs, early intervention, and social skills training. As a result, inaccurate diagnostic decisions for SLD yield two possible outcomes: (a) students who actually have SLD, but are deemed non-disabled, may be denied access to appropriate educational supports or (b) students who are inaccurately labeled as having SLD may be unintentionally placed at-risk for sharing the negative outcomes of students who have SLD. The identification of construct-irrelevant variance in norm-referenced IQ test score interpretations used to make SLD eligibility determinations may help reduce inappropriate diagnostic decisions for students suspected as having a SLD.

The importance of construct validity in norm-referenced test score interpretations used in the assessment of Specific Learning Disabilities. Norm-referenced scores are interpreted by comparing the performance of an examinee with the performance of a normative comparison group (Reynolds et al., 2006). Construct validity is compromised when the examinee and the normative comparison group do not share similar demographic characteristics (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999; APA, 2001). The cultural and linguistic makeup of the U.S. school-aged population is becoming increasingly diverse (Aud et al., 2010) and demographic characteristics such as culture and language diversity may serve as sources of construct-irrelevant variance (Hambleton & Li, 2005). IDEA requires that states have in place, "policies and procedures designed to prevent the inappropriate overidentification or disproportionate representation by race and ethnicity of children as children with disabilities" (§300.173). However, it is well documented that ethnic minority students and students who speak English as a second language (Ell) are disproportionately at risk for being identified as having SLD (U.S. Department of Education, Office of Special Education and Rehabilitative Services, Office of Special Education Programs, 2005; Zehler et al., 2003). This is especially true when definitions of SLD reduce the importance of general cognitive ability (Colarusso, Keel, & Dangel, 2001; McDermott, Goldberg, Watkins, Stanley, & Glutting, 2006). The presence of construct-irrelevant variance in norm-referenced IQ test score interpretations that are used to make SLD eligibility determinations may cause students who are culturally and linguistically diverse to be misidentified. Moreover, this may

help explain the disproportionate risk of some subgroups of the school-aged population for being identified as having SLD compared to others.

The Assessment of Specific Learning Disabilities Amongst Culturally and Linguistically Diverse Students

Assessment of students who are culturally and linguistically diverse requires special considerations. Professional standards for assessing individuals from diverse cultural and linguistic backgrounds have been outlined in numerous publications such as the Standards for Educational and Psychological Testing (2004) and the Code of Fair Testing Practices in Education (2004), as well as the APA's Guidelines for Test User Qualifications (2000) and Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations (1990). Assessment practices for cultural and linguistic minority students have also been addressed directly in the ethics codes of the APA (APA, 2002, Standards 9.03c & 9.06) and the National Association of School Psychologists (NASP; NASP, 2010, Section II.3.5). Lastly, IDEA provides legal mandates regarding general policies and practices used in the identification of educational disabilities for culturally and linguistically diverse students (IDEA, 2004, Title I, Part B §618d) as well as specific exclusionary criteria to be considered in SLD eligibility determinations (IDEA, 2004, Title I, Part B §300.309a).

The appropriate selection of test instruments and accurate interpretation of norm-referenced scores obtained from IQ tests are especially important in the assessment of culturally and linguistically diverse students (AERA, APA, & NCME, 1999; APA, 1990, Guideline 2d; APA, 2000, Standard 9.06; Joint Committee of Testing Practices, 2004, p. 3 & p. 9; NASP, 2010, Section II.3.5). Construct validity is only maintained when the examinee shares demographic characteristics with the normative comparison group (AERA, APA, & NCME, 1999). Thus, test selection must take into account examinee demographic characteristics, including culture and English language proficiency (AERA, APA, & NCME, 1999; APA, 2000; Joint Committee of Testing Practices, 2004, p. 3). In addition, IDEA requires that SLD eligibility determinations be made in the absence of cultural or linguistic bias (IDEA, 2004, Title I, Part B §300.309a). This means that learning difficulties that manifest from factors related to cultural and linguistic diversity cannot be learning *disabilities*. Lastly, psychologists have an ethical obligation to report any limitations to score interpretations that are derived from tests

with normative comparison groups in which the examinee is not adequately represented (APA, 2002, Standard 9.06; NASP, 2010, Section II.3.5). This information is necessary in order to facilitate effective communication between parents and educators about appropriate educational programming.

The examinee's level of English language proficiency yields additional considerations. Use of interpretation services is a practice endorsed by educational and psychological standards (AERA, APA, & NCME, 1999; APA, 1990, Guideline 6). The Standards state that, "when an adequately translated version of the test or a suitable nonverbal test is unavailable, assessment of individuals with limited proficiency in the language of the test should be conducted by a professionally trained bilingual examiner" (p. 95). Interpretation services are also used in the administration of psychological assessments. Psychologists are obligated to obtain informed consent from parental guardians and students to use interpretation services, take steps to ensure the confidentiality of the test results, and document (e.g., as in reports or recommendations) any limitations in test score interpretations that result from using an interpreter during an assessment (APA, 2002, Standard 9.03c). Deviation from these standards may inadvertently cause test scores to be misinterpreted and students to be inaccurately identified as having or not having SLD as a result of construct-irrelevant variance due to factors associated with differences between the cultural and linguistic makeup of the normative comparison group and the cultural and linguistic background of the examinee.

APA (1990) recommends that psychologists synthesize and document relevant cultural, linguistic, and sociopolitical factors to ensure that they are addressed in the assessment and intervention for psychological problems. Unfortunately, practitioners do not always follow professional standards, ethics, and regulations. Ochoa, Riccio, Jimenez, Garcia de Alba, and Sines (2004) surveyed 439 school psychologists' assessment practices with Ells. Approximately half of the survey respondents had some experience assessing these students and most reported using interpreters. However, one quarter of the interpreters were not trained in test administration. The continued use of contraindicated assessment practices concerns some researchers who believe that psychological training does not adequately prepare school psychologists for the assessment of students who are culturally and linguistically diverse (Ortiz & Dynda, 2005; Ortiz & Ochoa, 2005).

The Culture-Language Interpretive Matrix

Flanagan, Ortiz, and Alfonso (2007) created the Culture-Language Interpretive Matrix (C-LIM) to respond to these problems. The C-LIM is a 3 by 3 matrix that is hypothesized to help practitioners differentiate between individual students suspected of having a disability as either: (a) culturally and linguistically *dissimilar* to test score normative comparison groups for standardized intelligence tests or (b) culturally and linguistically *similar* to test score normative comparison groups for standardized intelligence tests. Flanagan et al. refer to this phenomenon as, "difference versus disorder" (Flanagan et al., 2007, p. 175). The vertical axis of the matrix represents *degree of cultural demand* and the horizontal axis of the matrix represents *degree of linguistic demand.* Contained within the nine cells are the subtests of specific standardized intelligence tests that were classified by Flanagan et al. as having *low*, *moderate*, or *high* cultural and linguistic demand. Classifications were based upon a review of the relevant literature, comparisons of the differences in the patterns of standardized test performance between monolingual and bilingual English-speakers, and expert consensus (Flanagan et al., 2007, p. 169). Each C-LIM corresponds to a specific intelligence test and Flanagan et al. created matrices to accompany a myriad of standardized intelligence tests. Figure 1 illustrates a generic C-LIM.

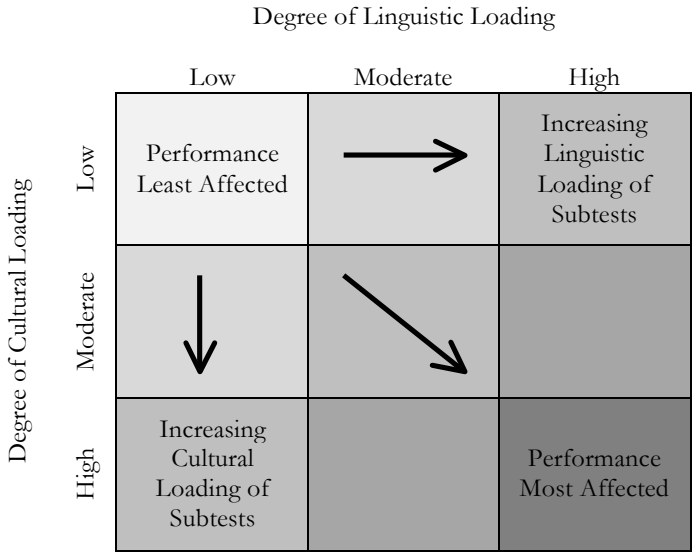

Figure 1. A generic Culture-Language Interpretive Matrix. Shaded areas represent the five levels of the degree of linguistic and cultural loading hypothesized to exist by Flanagan et al. (2007).

Interpretation of the C-LIM is conducted through three sequential steps. First, subtest scaled scores are converted into standardized scores with a mean of 100 and a standard deviation of 15. Second, the mean of each cell is computed. Lastly, the scores along the diagonal of the matrix are examined. Cell means that decline down the diagonal when the highest cell mean is located in the upper left-hand corner of the matrix and the lowest cell mean is located in the lower right-hand corner of the matrix are hypothesized to indicate difference and any other pattern of scores is hypothesized to indicate disorder (Flanagan et al., 2007, p. 178-181). Flanagan et al. offered the following interpretation of scores for students designated different, "practitioners must recognize that the invalidity of their results indicates that no interpretation can be made and no direct inferences drawn regarding levels of actual or true ability" (Flanagan et al., 2007, p. 197). Scores from students designated disordered do not conform to any predicted pattern and the source of these score patterns is hypothesized to be, "due to learning disability because the results will vary more a function of which test scores happen to be low and what constructs those tests are designed to measure" (Flanagan et al., 2007, p. 196-197).

The C-LIM is a new tool for use in the comprehensive evaluation of culturally and linguistically diverse students, but does it work as intended? Professional standards (AERA, APA, & NCME, 1999; APA, 2001; APA, 1990; Joint Committee of Testing Practices, 2004), ethics codes (APA, 2002; NASP, 2010), and educational laws (IDEA, 2004) require practitioners to select and interpret educational and psychological tests based upon the shared demographic characteristics of the examinee and the normative comparison group. Flanagan et al. (2007) claimed that the C-LIM, "may well prove to be of significant practical value in decreasing bias related to the selection and interpretation of tests" (p. 175).

Problems with the Culture-Language Test Classification system. An appraisal of the properties of the C-LIM as well as a review of the literature related to the diagnostic use of the C-LIM exposed several concerns. First, the categorization of subtests from tests of intelligence as having low, medium, or high cultural and linguistic demand within each C-LIM, also referred to as the Culture-Language Test Classifications (C-LTC), was determined in the absence of empirical evidence of reliability or validity. Flanagan et al. (2007) described the categorization procedure as involving

three steps. First, research on IQ test scores was reviewed and, "produced a rather strong consensus that bilinguals tended to perform about one standard deviation below the mean of monolinguals" (p. 169). This information alone is insufficient for identifying a diagnostic classification system because group mean differences do not adequately address individual differences. In fact, diagnostic groups are frequently heterogeneous (Garfield, 1978) and inferences drawn from group mean differences can be misleading as a result. The second and third classification procedures were described as follows:

> Second, data from many of these studies included mean scores for bilingual individuals on various tests, most commonly the Wechsler batteries. By aligning the tests in terms of mean differences, as compared to monolingual individuals, tests could be arranged in terms of the degree that performance was attenuated by cultural and linguistic differences. And third, because of the lack of research on bilinguals with many of the batteries currently in use, an expert consensus procedure was utilized to provide a logical basis for classifications of tests for which no existing data were available (Flanagan et al., p. 169).

Flanagan et al. (2007) classified the subtests from twenty intelligence tests as having low, medium, or high linguistic and cultural demand using these methods. Only five of the twenty C-LIMs are coordinated with Wechsler scales, which indicates that the remaining fifteen C-LIMs were solely classified based upon the "expert consensus procedure" (p. 169) described by Flanagan et al. This suggests that clinical judgment was the primary method used to classify IQ subtests as having low, moderate, or high linguistic and cultural demand. Problems arise when practitioners rely primarily on clinical judgment to make high stakes diagnostic decisions. It is well known that diagnostic decisions made within clinical settings that are based upon clinical judgment are less accurate than diagnostic decisions based upon actuarial judgment (Dawes & Corrigan, 1974; Dawes, Faust, & Meehl, 1989; Meehl, 1954). Consequently, the C-LIM subtest classification system has questionable reliability and validity.

Problems with empirical evidence gathered on the C-LIM. Empirical evidence regarding the validity of the C-LIM profile interpretations is sparse. Only five studies on the C-LIM have been conducted to date and only one of those five studies has been published in a peer-reviewed journal (Kranzler, Flores, & Coady, 2010). The remaining four studies consist of unpublished doctoral dissertations that were conducted under the direction of one of the C-LIM authors (Dhaniram-Beharry, 2008; Nieves-Brull, 2006; Tychanska, 2009; Verderosa, 2007).

Unpublished research on the previous edition of the C-LIM. The initial research on the C-LIM consists of four unpublished doctoral dissertations (Dhaniram-Beharry, 2008; Nieves-Brull, 2006; Tychanska, 2009; Verderosa, 2007). Two of the four dissertations were conducted on an earlier version of the C-LIM proposed by Flanagan and Ortiz (2001) (Nieves-Brull, 2006; Verderosa, 2007) and the other two dissertations were conducted on the current version of the C-LIM proposed by Flanagan et al. (2007) (Dhaniram-Beharry, 2008; Tychanska, 2009). Chronologically, the first dissertation was completed by Nieves-Brull (2006). Nieves-Brull (2006) sought to, "investigate the nature of the C-LTC and to provide an empirical investigation on the accuracy of the subtests from the WISC-III" (p. 15). Wechsler Intelligence Scale for Children-Third Edition (WISC-III) scores from a referred sample of 119 monolingual English speaking students (n = 53) and bilingual English/Spanish speaking students (n = 66) who were found ineligible for special education services by a multidisciplinary education team were used to empirically test several hypotheses.

First, Nieves-Brull (2006) tested the hypothesis that WISC-III subtest scores from the monolingual English speaking students and the bilingual English/Spanish speaking students would be similar to each other[1]. A Multivariate Analysis of Variance (MANOVA) was used to analyze the degree to which scores on the WISC-III standard battery subtests varied as a function of language status (i.e., monolingual English speakers vs. bilingual English/Spanish speakers). Main effects and post hoc tests indicated that scores from monolingual English speaking students were higher on average than scores from bilingual English/Spanish speaking students on 6 out of 10 WISC-III subtests. Flanagan et al. (2001) claimed that the C-LIM, "may well prove to be of significant practical value in decreasing bias related to the selection and interpretation of tests" (p. 250) through the interpretation of the different score pattern profile. However, no profile was investigated in the first analysis.

Next, Nieves-Brull (2006) used t tests to compare the degree to which WISC-III subtest scores from each language status group varied from three specified score pattern models. Model 1

[1] Flanagan et al. (2007) specify that standardized test score subtests must be converted into Standard Score units with a mean 100 and a Standard Deviation of 15. All research conducted on the Culture-Language Test Classification and Culture-Language Interpretive Matrix adhered to this practice. Study results are reported in Standard Score units unless otherwise specified from this point forward.

hypothesized that both language status groups' subtest scores would be statistically similar to 100. Nieves-Brull (2006) hypothesized that scores from the monolingual English speakers would be closer to 100 than scores from the bilingual English/Spanish speakers because, "the C-LTC/C-LIM predicts that the monolingual group will score at or near the mean for the test (SS = 100), whereas the bilingual group will score lower in general because of attenuation on the verbally loaded tests" (p. 16). Model 2 hypothesized that both language status groups' subtest scores would be statistically similar to 85. Rationale for this score pattern was described as testing, "the notion that the sample contained individuals who somehow lack ability in general, irrespective of whether they were bilingual or monolingual" (p. 16). Nieves-Brull (2006) also hypothesized that the scores from the bilingual English/Spanish speakers would be closer to 85 than scores from the monolingual English speakers because of, "attenuation of the verbally loaded scores" (p. 16). Lastly, model 3 hypothesized that both language status groups' subtest scores would be statistically similar to scores from a pattern that included an average score of 85 on the Information, Similarities, Vocabulary, and Comprehension subtests, an average score of 92 on the Arithmetic and Picture Completion subtests, and an average score of 95 on the Coding, Picture Arrangement, Block Design, and Object Assembly subtests. Nieves-Brull (2006) hypothesized that, "the [average] score obtained by the bilingual group will be closer to the predicted pattern that demonstrates a systematic decline in performance as the cultural loading and linguistic demands of the subtests increase" (p. 17). No further rationale were provided for the score patterns designated in each model.

Results for the analysis conducted on model 1 revealed no significant differences in the degree to which WISC-III subtest scores varied from the designated value of 100 between each language status group. Nieves-Brull (2006) interpreted these findings to mean, "the model was a better fit for the monolinguals than the bilinguals" (p. 30) because the scores from the bilingual English/Spanish speaking participants varied more from 100 than the scores from the monolingual English speaking participants. The results of the analysis do not support Nieves-Brull's interpretation. Furthermore, the difference between WISC-III subtest scores from the monolingual English speaking participants and 100 was only 0.75 standard score points higher than the difference between WISC-III

subtest scores from bilingual English/Spanish speaking participants. A difference of this magnitude is of no practical importance.

Interpretation of results from analyses conducted on the remaining two models was also problematic. Analyses revealed that WISC-III subtest scores from bilingual English/Spanish speakers varied significantly less from 85 than WISC-III subtest scores from monolingual English speakers. Additionally, WISC-III subtest scores from bilingual English/Spanish speakers varied less from the score pattern hypothesized in the model 3 condition compared to the WISC-III subtest scores from monolingual English speakers. This indicates that, on average, bilingual English/Spanish speakers' WISC-III subtest scores were closer to 85 and followed the pattern of decline hypothesized by Flanagan et al. (2001) to exist for students who are culturally and linguistically different than the WISC-III normative comparison group more so than monolingual English speakers' WISC-III subtest scores. However, these results do not indicate the frequency with which individual bilingual English/Spanish speakers' WISC-III subtest scores followed Flanagan et al.'s hypothesized pattern of decline. It is well known that mean differences do not translate to individual differences (Elwood, 1993) and individual statistical information is missing from this analysis.

Lastly, Nieves-Brull (2006) made an attempt to provide individual diagnostic information by calculating the, "frequency with which students' scores reached the smallest mean squared difference by model" (p. 41). Results indicated that scores from only 24% of the bilingual English/Spanish speakers in her sample reached the, "smallest mean squared difference" in the model 3 condition. This means that at least 76% of the scores from her sample of non-disabled bilingual English/Spanish speakers varied more from Flanagan et al.'s (2001) hypothesized score pattern of decline for students who are culturally and linguistically diverse from the WISC-III normative comparison group than the score patterns designated in the other two model conditions. These frequencies suggest that the C-LIM may be inaccurate at making individual diagnostic decisions. In spite of these findings, Nieves-Brull (2006) interpreted the results of her study to suggest, "that the C-LTC can assist school psychologists in making a distinction between students with learning disorders and students with linguistic differences" (p. 48). It is unclear how that conclusion was drawn from the study results.

Following chronological order, Verderosa (2007) was the next dissertation completed on the C-LIM. This study was also conducted on the previous version of the matrix. Verderosa (2007) examined scores on the Differential Ability Scales (DAS; Elliot, 1990) from a referred sample of 60 bilingual preschoolers in order to, "examine the validity of the DAS matrix classifications with a bilingual preschool population" (p. 21) and to investigate, "the utility of using the DAS when assessing bilingual preschoolers" (p. 21). To address these goals, Verderosa (2007) used *t* tests to empirically test the existence of statistically significant differences between cell means of the DAS C-LIM hypothesized by Flanagan and Ortiz (2001) to be affected by linguistic and cultural differences to a low, moderate, and high degree. For example, cell means that contained subtests deemed to have low linguistic and cultural loading were hypothesized by Flanagan and Ortiz (2001) to be greater than cell means that contained subtests deemed to have moderate and/or high linguistic and cultural loading. It should also be noted that the reader must assume that participants included students who were found ineligible for special education services because detailed demographic information was not provided.

Results of hypothesis testing indicated significant differences between cell means in a pattern that matched Flanagan and Oritz's (2001) hypothesized DAS C-LIM with one exception. The mean score for the low linguistic demand/moderate cultural demand cell was not significantly greater than the mean score for the low linguistic demand/high cultural demand cell. This suggests that, on average, scores from Verderosa's (2007) sample of Ells varied somewhat from the pattern of scores hypothesized by Flanagan and Ortiz (2001). In addition, no individual statistics were calculated to determine the frequency of participants' scores that followed the predicted pattern. Information about the diagnostic utility of the DAS C-LIM cannot be inferred from this analysis as a result, despite Verderosa's (2007) aim to examine, "the utility of using the DAS when assessing bilingual preschoolers" (p. 21).

Verderosa (2007) also discussed, "findings of exploratory analyses conducted" (p. 29). Results of these analyses indicated that scores from Verderosa's (2007) sample of Ells were significantly greater on the DAS Nonverbal Cluster than on the Verbal Cluster. Verderosa (2007) examined this pattern in greater detail by conducting an ANOVA to determine how subtest and broad

cluster scores on the DAS varied as a function of participants' language category reported on a Home Language Survey. Results of the ANOVA indicated that scores from participants who were designated English language dominant at home were significantly greater than scores from participants who were designated equal English and Spanish language dominant at home, Spanish language dominant at home, and/or as speaking Spanish only at home on all DAS subtests with the exception of Verbal Comprehension. Scores from participants reported to speak predominantly Spanish at home were significantly higher on the Verbal Comprehension subtest compared to scores from participants who were determined to speak predominantly English at home, both English and Spanish equally, and only Spanish at home. This pattern does not support Flanagan and Ortiz's (2001) hypotheses that, "individuals whose linguistic backgrounds are at significant variance with the linguistic backgrounds of individuals who comprise the norm group…will tend to score lower on standardized tests as a function of their reliance on language." (p. 246).

An examination of the degree to which DAS broad cluster scores varied as a function of participants' language category reported on a Home Language Survey yielded a similar pattern. Scores from participants whose home language was designated predominantly English were significantly greater than scores from participants whose home language was reported to include English and Spanish equally, predominantly Spanish, and only Spanish for the General Cognitive Cluster and the Nonverbal Cluster. However, scores from participants whose home language was reported as predominately Spanish were significantly higher than scores from participants whose home language was designated to be predominantly English, both English and Spanish, and only Spanish on the Verbal Cluster. Once again, patterns of scores emerged that did not support Flanagan and Oritz's (2001) hypothesized score pattern regarding, "the amount of linguistic skill required by the various tests of intelligence and cognitive ability" (p. 245).

Lastly, a similar analysis was conducted to determine the degree to which DAS scores varied as a function of culture as measured by the Bidimensional Acculturation Scale (BAS; Marin & Gamba, 1996). No statistical differences emerged between participants grouped by acculturation status on any DAS subtests or broad cluster scores. Again, these results did not support Flanagan and Ortiz's (2001) hypothesis that, "tests that are more process oriented and that contain more novel, culture-

reduced stimuli and communicative requirements might yield scores that are fairer estimates of ability or skill" because, "they would be less subject to attenuating influences from an individual's level of exposure to mainstream culture" (p. 245). The empirical evidence gathered in Verderosa (2007) does not support Flanagan and Oritz's (2001) hypotheses surrounding the score patterns delineated in the DAS C-LIM. Nevertheless, Verderosa (2007) concluded that, "results of this investigation provide support for Flanagan and Ortiz's (2001) matrix system" (p. 67). Again, it is unclear how this conclusion was drawn.

Unpublished research on the current edition of the C-LIM. Dhaniram-Beharry (2008) was the first dissertation completed on the current version of the C-LIM (Flanagan et al., 2007). It was also the only study that failed to include a participant group of Ells. Dhaniram-Beharry (2008) attempted to, "investigate the validity of the patterns of performance predicted by the C-LTC and C-LIM" (p. 14) by determining, "if a systematic pattern of declining test scores emerges as a function of disability, gender, grade level, and race" (p. 13). Specifically, Flanagan et al. hypothesized that scores that follow a pattern of decline down the diagonal of the C-LIM, "suggest that test results were influenced primarily by level of acculturation and limited English proficiency rather than by actual ability" (p. 181). Therefore, Dhaniram-Beharry (2008) asserted that, "this pattern [of decline] should only occur with a group of English language learners since the matrix was designed to take into consideration their level of language and culture and how these factors can attenuate test performance" (p. 14).

To achieve this goal, WISC-III, Wechsler Intelligence Scales for Children-Fourth Edition (WISC-IV), and/or Woodcock-Johnson Tests of Cognitive Abilities-Third Edition (WJ-III Cog) scores from a group of 64 English-speaking students who were found eligible for special education services as students with SLD by a multidisciplinary education team were collapsed into a single C-LIM. For example, subtests from all three IQ tests hypothesized by Flanagan et al. (2007) to contain low linguistic demand/low cultural demand were placed in the low linguistic demand/low cultural demand cell within the matrix, subtests from all three IQ tests hypothesized to contain moderate linguistic demand/low cultural demand were placed in the moderate linguistic demand/low cultural demand cell within the matrix, and so forth. Scores from each of the four participant subgroups (i.e.,

formed according to disability, gender, grade level, and race) were statistically compared to a pattern of scores hypothesized by Flanagan et al. to exist for students determined to be *slightly* culturally and linguistically different from the normative comparison group of any given IQ test using *t* tests. The score pattern used to designate slightly different in Dhaniram-Beharry's (2008) study consisted of: (a) a cell mean of 96 in the level 1 cell within the C-LIM (i.e., low degree of linguistic demand/low degree of cultural demand), (b) a cell mean of 94 in the level 2 cells within the C-LIM (i.e., moderate degree of linguistic demand/low degree of cultural demand and low degree of linguistic demand/moderate degree of cultural demand), (c) a cell mean of 91 in the level 3 cells within the C-LIM (i.e., high degree of linguistic demand/low degree of cultural demand, moderate degree of linguistic demand/moderate degree of cultural demand, and low degree of linguistic demand/high degree of cultural demand cells), (d) a cell mean of 88 in the level 4 cells within the C-LIM (i.e., high degree of linguistic demand/moderate degree of cultural demand and moderate degree of linguistic demand/high degree of cultural demand cells) and (e) a cell mean of 83 in the level 5 cells within the C-LIM (i.e., high degree of linguistic demand/high degree of cultural demand).

It should be noted that Flanagan et al. (2007) provided a range of scores that are hypothesized to exist for students designated slightly culturally and linguistically different than the normative group of a given IQ test. For example, Flanagan et al. stated that cell means from students who are designated slightly different will be 3-5 points lower than the cells means from the normative group in the low linguistic demand/low cultural demand cell. Dhaniram-Beharry (2008) chose to use a numerical value that was 4 points lower than the normative group to indicate slightly different in her study. This value falls within the 3-5 point range; however, it does not cover the entire range of scores hypothesized by Flanagan and Ortiz (2001) to indicate that a student is slightly culturally and linguistically different from the normative group.

Dhaniram-Beharry's (2008) first hypothesis was that, "a systematic decline in the scores of the LD students will emerge mimicking the pattern of predicted performance of CLD students" (p. 14). Results from a series of *t* tests indicated that, on average, sample scores were significantly higher than Flanagan et al.'s (2007) predicted scores on 2 out of the 5 cell levels within the combined WISC-III, WISC-IV, and WJ-III Cog C-LIM. This means that subtest scores from Dhaniram-Beharry's

(2008) sample of English speaking students with SLD were statistically similar to Flanagan et al.'s hypothesized score pattern for students who are culturally and linguistically different from the normative group of the three IQ tests on 3 out of the 5 cell levels. Nevertheless, examining the statistical mean differences between participants' scores and scores predicted by Flanagan et al. on the C-LIM on cell levels does not address the hypothesis as to whether participant scores will mimic, "the pattern of predicted performance of CLD students" (p. 14). Score patterns were not investigated in any analysis.

These same inadequate statistical methods were used to test the remaining hypotheses that, "a systematic decline in the scores" for the rest of the subgroups of participants' would, "emerge mimicking the pattern of predicted performance of CLD students" (p. 14). Results indicated that scores from participants identified as Caucasian were statistically different from Flanagan et al.'s (2007) predicted scores on 1 out of the 5 cell levels, scores from participants in grades 4 through 8 were statistically different from Flanagan et al.'s predicted scores on 3 out of the 5 cell levels, and scores from the remaining participant subgroups were statistically different from Flanagan et al.'s predicted scores on 2 out of the 5 cell levels. No meaningful information regarding the validity of the C-LIM interpretations can be derived from these results.

Of additional concern, Dhaniram-Beharry (2008) stated that, "some students were not administered subtests classified at level one, and subtests for level four of the Culture-Language-Interpretive do not exist on the Wechsler scales" (p. 33) so scores from "levels one and four of the matrix were analyzed using only WJ-III data, whereas at levels two, three and five analyses were conducted based on cognitive scores from both test batteries" (p. 33). This suggests that different samples were used in each analysis, which further complicates the interpretation of Dhaniram-Beharry's (2008) results. Notwithstanding, combining scores from standardized IQ tests using cross-battery methods have been widely criticized on various points including comparing subtest scores from instruments with different normative samples, administering subtests in a non-standardized presentation order, and the lack of established external validity for such practices (Floyd, Clark, & Shadish, 2008; Glutting, Watkins, & Youngstrom, 2003).

Lastly, Dhaniram-Beharry (2008) attempted to calculate the Euclidean distance of each individual participant's score from the score predicted by Flanagan et al. (2007). Dhaniram-Beharry (2008) used the Expectation Maximization Algorithm to estimate missing data in order to create, "a complete data set, such that all 64 cases had scores at every matrix level" (p. 35). This is a curious proposition given that, "subtests for level four of the Culture-Language-Interpretive [Matrix] do not exist on the Wechsler scales" (p. 33). Estimating the missing data from the, "students [who] were not administered subtests classified at level one" (p. 33) is a viable option for estimating values for missing data (Baraldi & Enders, 2010) at level one because those values were truly missing. However, estimated values for "missing" data at level four represent values for subtests that "do not exist," which complicates the interpretation of those results.

Dhaniram-Beharry (2008) suggested her results indicate that, "school psychologists can use the C-LTC and C-LIM to guide them when interpreting test data of CLD [culturally and linguistically diverse] students" (p. 49) in spite of these methodological problems. One final consideration in the interpretation of the results of Dhaniram-Beharry (2008) is the absence of a sample of Ells. How can the results be interpreted in this fashion when score patterns for culturally and linguistically diverse students were not examined?

The final study conducted on the current C-LIM was completed by Tychanska (2009). The purpose of Tychanska (2009) was to, "determine whether distinctive patterns of performance exist as a function of language status and disability type, and whether such patterns may be derived using the C-LIM" (p. 20) because Flanagan et al. (2007) stated that Ells with a, "true speech-language impairment" may "show a clear, declining pattern" (p. 198). Participants included a referred sample of 182 students who were administered the WISC-III, WISC-IV, or Wechsler Preschool and Primary Scales of Intelligence-Third Edition (WPSSI-III) as part of a special education evaluation. Once again, scores from all three IQ tests were collapsed into a single C-LIM. The following participant subgroups were formed for the purpose of the study: (a) English speakers who were identified as having SLD by a multidisciplinary evaluation team (n = 50), (b) English speakers who were identified as having only Speech and Language disabilities by a multidisciplinary evaluation team (n = 31), (c) Ells who were identified as having SLD by a multidisciplinary evaluation team (n = 65), and (d) Ells

who were identified as having only Speech and Language disabilities by a multidisciplinary evaluation team (n = 36). Students with no identified disabilities were not included in the study because, "their patterns of performance (mostly average across the board) already represent the basis of normative functioning" and "there is no reason to repeat such testing" (p. 30).

Tychanska (2009) first hypothesized that scores from Ells with identified Speech and Language disabilities would exhibit a, "systematic pattern of decline in test performance" that is "more rapid and severe than the typical pattern expected for non-disabled ELs [English language learners] (those without any disorder)" (p. 29). Additionally, Tychanska (2009) hypothesized that, "the drop in performance will be more marked and pronounced on the linguistically demanding tests more so than on the culturally loaded ones" (p. 29). To empirically test these hypotheses, Tychanska (2009) conducted t tests between the mean subtest scores of Ells with SLD and 100 (to designate the normative test mean) as well as t tests between the mean subtest scores of Ells with Speech and Language disabilities and 100. Results indicated that mean subtest scores from Ells with SLD were significantly lower than 100 on 14 of the 15 subtests included in the C-LIM from the WISC-III, WISC-IV, and WPSSI-III and that mean subtest scores from Ells with Speech and Language disabilities were significantly lower than 100 on 8 of the 15 subtests included in the study.

Results from mean difference testing on individual subtests does not address whether a "systematic pattern of decline in test performance" (p. 29) exists. However, these same inappropriate statistical methods were used to test the second hypothesis that scores from English speaking participants with identified Speech and Language disabilities would exhibit, "a systematic pattern of decline in test performance" (p. 31) when compared to, "the performance of non-disabled English speaking peers" (p. 31) and that, "the drop in performance will be marked and pronounced due primarily to the linguistic demand of the tests and less so to the cultural loading of the tests" (p. 31). Results indicated that mean subtest scores from English speaking participants with SLD were significantly lower than 100 on 5 out of the 15 subtests in the combined study C-LIM and that mean subtest scores from English speaking participants identified as having Speech and Language disabilities were significantly lower than 100 on 9 out of the 15 subtests in the combined study C-LIM. Meaningful information regarding the degree to which scores from participant subgroups

followed Flanagan et al.'s (2007) hypothesized different profile was not gained from these analyses. Moreover, these results are not surprising given that it is not uncommon for referral samples to have lower mean subtest scores compared to the normative comparison group (Canivez & Watkins, 1998; Watkins, 2010).

Tychanska (2009) also conducted *t* tests between individual C-LIM cell means and 100 to determine the existence of significant differences for each participant subgroup, separately. Results indicated that mean cell scores from English speaking participants with SLD were significantly lower than 100 on 3 out of the 9 cells within the combined study C-LIM and mean cell scores from English speaking participants with Speech and Language disabilities were significantly lower than 100 on 4 out of the 9 cells within the combined study C-LIM. Additionally, results indicated that mean cell scores from participants who were designated Ells with SLD were significantly lower than 100 on all 9 of the cells within the combined study C-LIM and mean scores from Ells with Speech and Language disabilities were significantly lower than 100 on 4 out of the 9 cells within the combined study C-LIM. Again, these results are unsurprising given that they came from a referred sample (Canivez & Watkins, 1998; Watkins, 2010) and they do not address the degree to which score patterns followed Flanagan et al.'s (2007) hypothesized different score pattern profile.

Fortunately, C-LIM cell means were reported for each participant subgroup. Flanagan et al. (2007) claimed that Ells with "true speech-language impairment" may "show a clear, declining pattern" (p. 198). Scores from Tychanska's (2009) sample of Ells with Speech and Language disabilities did not follow Flanagan et al.'s different score pattern profile. However, mean scores from Ells with SLD and mean scores from English speakers with Speech and Language disabilities did follow Flanagan et al.'s different score pattern profile perfectly. These results further necessitate the evaluation of C-LIM profile interpretations for individual students.

An explanation for why scores from the participant group comprised of English speakers with Speech and Language disabilities followed Flanagan et al.'s (2007) different score profile was not discussed. However, Tychanska (2009) provided the following circular explanation for the emergence of the declining pattern of scores from the participants designated Ells with SLD, "it demonstrates a systematic influence of cultural and linguistic variables and, therefore, does not support the

expectation that this is a group that is identified correctly as LD" (p. 77). Tychanska (2009) elaborated by stating that the, "finding of a lower mean supports the notion that the English learner LD group was probably not LD at all, but it may have been misdiagnosed in the first place" (p. 77). Once more, it is not uncommon for referred samples to have lower mean scores than the normative comparison group of a given IQ test (Canivez & Watkins, 1998; Watkins, 2010) and lower mean subtest scores have no bearing on whether or not individuals are accurately diagnosed as having SLD.

Published research on the current version of the C-LIM. Kranzler, Flores, and Coady (2010) conducted the only published study on the C-LIM. Kranzler et al. administered the WJ-III Cog to a non-referred sample of 46 students receiving English as a second language services to, "examine the predicted effects of cultural loading and linguistic demand on test performance" (p. 435). To achieve this goal, the authors empirically tested the degree to which statistically significant differences existed between matrix cells that contained scores categorized within the C-LTC as having low, moderate, and high linguistic and/or cultural demand across rows (i.e., Degree of Linguistic Demand), columns (i.e., Degree of Cultural Demand), and the diagonal (i.e., Degree of Cultural and Linguistic Demand, Combined) of the WJ-III Cog C-LIM. For example, Flanagan et al. (2007) hypothesized that cell means across the rows of the matrix would decrease from left to right and that cell means down the columns of the matrix would decrease from top to bottom.

Multiple analyses were conducted to test Flanagan et al.'s (2007) hypotheses. First, Kranzler et al. (2010) calculated the Pearson product-moment correlation between participant scores on an English language proficiency exam and scores on the WJ-III Cog (e.g., both subtest and General Intellectual Ability scores were used in the analysis). All correlations were non-significant, which suggests a weak relationship between the scores on these two tests. These results do not support Flanagan et al.'s hypothesis that, "in cases in which individuals are limited in English proficiency or; for whatever reasons, are not developmentally equivalent in language proficiency to the norm group, the result will be bias" (p. 158). Second, within-subjects Analysis of Variances (ANOVA) and *t* tests were used to test whether differences between cells that contained subtests classified as low, moderate, and high linguistic and/or cultural demand were statistically significant across rows, columns, and the diagonal of the matrix. Main effects were only significant for differences detected

across the diagonal of the matrix (i.e., Degree of Cultural and Linguistic Demand, Combined), the third column of the matrix (i.e., high Degree of Linguistic Demand), and the second row of the matrix (i.e., moderate Degree of Cultural Demand). However, post hoc tests nullified the clinical importance of the main effects. Post hoc tests revealed non-significant differences between individual cells or significant differences in a pattern opposite to the hypothesis of Flanagan et al. for all three main effects. These results also do not support Flanagan et al.'s hypotheses that, "cultural and linguistic differences serve to artificially depress the scores of diverse individuals" (p. 176).

Some diagnostic utility information was also collected in this study. Kranzler et al. (2010) calculated the percent of their sample that followed the declining pattern of cell means across the diagonal of the matrix hypothesized by Flanagan et al. (2007) to exist for students who are culturally and linguistically different from IQ test normative comparison groups. The study participants' mean score pattern and the frequency of the existence of this score pattern for individual study participants were both reported. As a group, mean scores of participants followed the hypothesized score pattern. However, scores from only 37% of individual study participants exhibited the hypothesized score pattern, whereas 100% of study participants were students receiving English as a second language services with no identified learning disabilities. This means that 63% of the study sample would be inaccurately determined to be similar to the WJ-III Cog normative comparison group and potentially diagnosed as having SLD according to Flanagan et al.'s C-LIM framework. Flipping a coin would be more accurate according to these statistics. Consequently, the authors concluded that the results of their study did not, "substantiate the use of C-LIMs for the assessment of cognitive abilities for children and youth from diverse backgrounds" (p. 443).

Summary of research on the C-LIM. To summarize, Flanagan et al. (2007) hypothesized that scores from students who are culturally and linguistically different than the normative comparison group of a given IQ test would follow a pattern of decline (e.g., designated different) and that scores from students who are culturally and linguistically similar to the normative comparison group of a given IQ test and who possibly have SLD would not follow any specific pattern (e.g., designated disordered). However, only one study to date has identified this hypothesized different profile amongst the scores of a sample of students who are identified as culturally and linguistically

diverse. Kranzler et al. (2010) reported that mean scores from their sample of students receiving English as a second language services exhibited Flanagan et al.'s different profile for the WJ-III Cog, but they also reported that only 37% of individual participant scores followed the different profile, whereas 100% of their participants were culturally and linguistically different from the WJ-III Cog normative group. Furthermore, Kranzler et al. was the only study that reported results in support of the discriminate validity of the C-LIM. Verderosa (2007) found that mean scores from her sample of bilingual preschoolers on the DAS did not follow Flanagan et al.'s predicted different profile and results from Tychanska (2009) indicated that scores from her sample of English speakers with Speech and Language disabilities and scores from Ells with SLD matched Flanagan et al.'s different profile on the Wechsler scales, which does not support Flanagan et al.'s hypotheses. Neither Dhaniram-Beharry (2008) nor Nieves-Brull (2006) attempted to calculate the degree to which mean scores or individual scores from participants matched the hypothesized different profile. This suggests that there is more evidence that the C-LIM cannot discriminate between various subgroups of the school-aged population than there is evidence in support of the C-LIM classification system. Lastly, Kranzler et al. was the only study that reported the frequency with which individual scores from culturally and linguistically diverse students actually followed Flanagan et al.'s hypothesized pattern of decline, which is a curious anomaly given the ease of such a calculation.

Additionally, the statistical methods used to investigate the C-LIM in these studies did not address the degree to which the C-LIM produces accurate diagnostic decisions. Rather, mean difference tests were used. The existence of *mean* differences is insufficient for identifying the degree to which *individual* differences exist because score distributions between clinical and non-clinical samples overlap (Metz, 1978; Swets, Dawes, & Monahan, 2000). Figure 2 represents the hypothetical score distributions for a clinical and non-clinical sample.

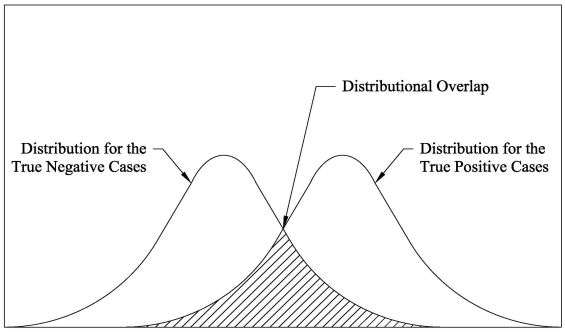

Figure 2. Hypothetical score distributions for a clinical and non-clinical sample.

Need for diagnostic utility statistics. The computation of diagnostic utility statistics is required to make judgments about the accuracy of a diagnostic tool. These statistics have been used extensively for the validation of psychological tests (Elwood, 1993; Gerardi, Keane, & Penk, 1989; Moldin, Gottesman, Rice, & Erlenmeyer-Kimling, 1991; Olin, Schneider, Eaton, Zemansky, & Pollock, 1992; Rapp, Parisi, Walsh, & Wallace, 1988) as well as for the validation of test score interpretations such as in profile analysis (see Watkins, 2003 for a comprehensive summary of the literature). Binary outcome data is a prerequisite for calculating these statistical parameters (Swets et al., 2000). Fortunately, the C-LIM produces a binary outcome as a result of the cut-score proposed by Flanagan et al.'s (2007) hypothesized different profile. Test scores either follow a hypothesized pattern of decline (e.g., different) or they do not (e.g., disordered). This "yes-no" decision inevitably produces both correct and incorrect diagnostic decisions because no test is 100% accurate.

Individuals can be characterized as either having a trait or not having a trait in the true state of the world. Therefore, four types of decisions are made by diagnostic tests: (a) true positive decisions in which individuals receive a positive diagnosis and they truly have the trait, (b) true negative decisions in which individuals receive a negative diagnosis and they truly do not have the trait, (c) false positive decisions in which individuals receive a positive diagnosis, but they do not truly have the trait, and (d) false negative decisions in which individuals receive a negative diagnosis, but they truly have the trait. Diagnostic utility statistics quantitatively describe the relationship of these decisions with the true states of specified traits (Cohen, 1990; Kessel & Zimmerman, 1993; Metz,

1978; Rosnow & Rosenthal, 1989). Sensitivity describes the probability of individuals who received a positive diagnosis from a particular diagnostic test, given that the trait is present. Specificity describes the probability of individuals who received a negative diagnosis from a particular diagnostic test, given that the trait is absent. The positive predictive power of a diagnostic test refers to the probability that a trait is present, given a positive test result. Conversely, the negative predictive power of a diagnostic test refers to the probability that a trait is absent, given a negative test result (Fawcett, 2006; Metz, 1978). This methodology can be used to provide information about the accuracy of the cut-score hypothesized by Flanagan et al. (2007).

Unfortunately, these diagnostic utility statistics are influenced by the base rate of a given trait (Elwood, 1993; Meehl & Rosen, 1955; Swets et al., 2000). For example, the chances of the C-LIM accurately identifying a student as culturally and linguistically different from the normative comparison group of a given IQ test are greater in a sample where there is a higher concentration of students who are culturally and linguistically different from the normative comparison group of a given IQ test than in a sample where most students share demographic characteristics with a given IQ test normative comparison group. Samples taken from different geographical regions with different base rates may yield different statistical parameters as a result. The receiver operating characteristic (ROC) curve is a statistical technique that eliminates the influence of base rates on diagnostic utility statistics to determine the accuracy of tests that are designed to discriminate between groups of individuals (Metz, 1978; Pintea & Moldovan, 2009). This is achieved by plotting 1 - specificity on the x-axis of a graph and sensitivity on the y-axis of the same graph for all possible cut-scores of a diagnostic test (Meehl & Rosen, 1955; Metz, 1978; Swets et al., 2000). The ensuing graph illustrates the relationship between the sensitivity and specificity of a given test for the full range of possible cut-score values (Fawcett, 2006), which gives a complete description of the diagnostic performance of the test (Pepe, 2003). A curve can then be fitted to the plotted points and the Area Under the Curve (AUC) can be used to interpret the overall accuracy of the diagnostic test. The ROC curve will move towards the upper left-hand corner of the graph as the accuracy of a diagnostic test increases. Figure 3 illustrates a graph with ROC curves at various AUC values.

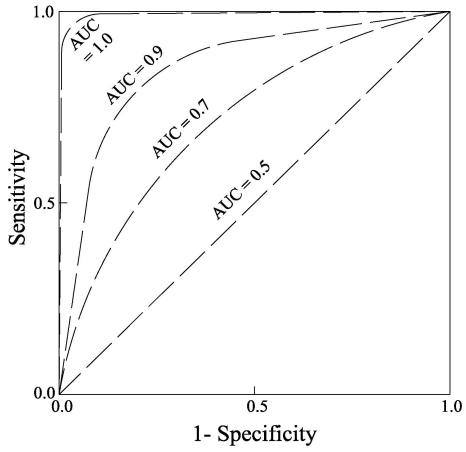

Figure 3. Hypothetical ROC curve and various AUC values.

AUC values have a simple interpretation. Assume that one person is randomly selected from a group of students who are culturally and linguistically different from the normative comparison group of a given IQ test and one person is randomly selected from a group of students contained in the normative comparison group of that particular test. The AUC value is the probability that the person chosen from the group of students who are culturally and linguistically different from the normative comparison group of the given IQ test will be correctly identified by the test (i.e., the C-LIM in this example). For instance, an AUC value of .50 would mean that the diagnostic test is no more accurate than chance (Hanley & McNeil, 1982; Shapiro, 1999). This means that the test in question has no ability to discriminate between the group of students who are culturally and linguistically different from the normative comparison group of the IQ test in question and the group of students contained in the IQ test's normative comparison group. Swets (1988) and Streiner and Cairney (2007) provided the following anchors for other AUC values: (a) AUC values between .50 and .70 characterize low accuracy, (b) AUC values between .70 and .90 characterize medium accuracy, and (c) AUC values between .90 and 1.00 characterize high accuracy.

Flanagan et al. (2007) asserted that the C-LIM, "may well prove to be of significant practical value in decreasing bias related to the selection and interpretation of tests" (p. 175). However, only one study to date has found that the different profile exists among the mean scores of students who are culturally and linguistically different from IQ test score normative comparison groups (Kranzler et al. 2010) and no study to date has found that this hypothesized score profile overwhelmingly exists

among culturally and linguistically diverse students at the individual diagnostic level (Dhaniram-Beharry, 2008; Kranzler et al., 2010; Nieves-Brull, 2006; Tychanska, 2009; Verderosa, 2007). Evidence in support of Flanagan et al.'s claim resides solely in the interpretations of the results of unpublished dissertations that relied on mean difference testing of individual cells or groups of cells to draw conclusions – none of which reported the degree to which individual sample scores followed Flanagan et al.'s hypothesized score profiles (Dhaniram-Beharry, 2008; Nieves-Brull, 2006; Tychanska, 2009; Verderosa, 2007).

Nevertheless, significant mean differences may suggest that the C-LIM can demonstrate discriminate validity, but they do not yield individual diagnostic accuracy (Watkins et al., 2005). High-stakes individual diagnostic decisions, such as those regarding SLD identification, have dire consequences for students both short- and long-term (e.g., Bruck, 1987; Finucci, 1986; Finucci et al., 1984; Kurzweil, 1992; Schonhaut & Satz, 1983; Sitlington & Frank, 1990; Spreen, 1988; Wagner et al., 2005) and it would be unethical to encourage practitioners to use a tool with questionable or unknown reliability, validity, and clinical utility characteristics to make such high-stakes decisions. Furthermore, standards 4.20 and 12.6 outlined in The Standards of Educational and Psychological Testing (1999) state, "when feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria" (p. 60) and specify that practitioners should choose "a test for which there is evidence of the test's ability to distinguish between the two or more diagnostic groups of concern rather than merely to distinguish abnormal cases from the general population" (p. 132) when differential diagnosis is warranted. Therefore, the purpose of the present study was to investigate the degree to which the C-LIM accurately discriminates between participants who are: (a) Ells with SLD, (b) Ells without SLD and (c) monolingual English speakers without a disabling condition using diagnostic utility statistics. It was hypothesized that the C-LIM cannot meaningfully distinguish between these groups of students given the results of current research.

Chapter 2

Method

Participants

Two main groups of participants were included in the present study: (a) Ells and (b) monolingual English speakers. The sample of Ells included 86 school-aged children (56 males, 30 females) aged 6 to 16 years (M = 11.3, SD = 2.4) who were referred for a special education evaluation to determine initial or continued eligibility as students with an educational disability in one of two school districts in the greater Phoenix metropolitan area. Participants from the referred sample were further disaggregated by the presence of SLD according to the four most commonly used criteria in applied practice (Mercer et al., 1996; Reschly & Hosp, 2004). Detailed demographic information for the referred participant subgroups is located in table 1.

The sample of monolingual English speakers without a disabling condition was extracted from the WISC-IV normative sample. Permission was granted from National Computer Systems Pearson, Incorporated to use the WISC-IV normative sample in the present study[2]. Participants included 2,033 students (1,004 males, 1,029 females) aged 6 to 16 years old (M = 11.5, SD = 3.2) who were administered the WISC-IV during the normative phase of test development. None of the participants were diagnosed as having educational related disabilities, but 19 (0.9%) of the participants were diagnosed as gifted and talented by independent practitioners.

[2] Standardization data from the *Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV)*.

Table 1

Demographic Information for 86 Ells Tested for Special Education Eligibility Disaggregated by SLD Criteria

Demographic characteristic	MET decision		Regression discrepancy		One standard deviation discrepancy		Two standard deviation discrepancy	
	SLD	No SLD	SLD	No SLD	SLD	No SLD	SLD	No SLD
n	76	10	55	31	48	38	13	73
Age in years								
Minimum	6.3	6.8	6.3	6.8	6.3	6.8	6.3	6.8
Maximum	16.2	16.4	14.7	16.4	14.7	16.4	14.2	16.4
M	11.5	10.3	11.3	11.7	11.1	11.7	11.1	11.4
SD	2.3	3	2.4	2.5	2.5	2.4	2.5	2.4
Gender *n*(%)								
Males	51(67.1%)	5(50%)	39(71%)	17(54.8%)	33(68.8%)	23(60.5%)	9(69.2%)	47(64.4%)
Females	25(32.9%)	5(50%)	16(29%)	14(45.2%)	15(31.3%)	15(39.5%)	4(30.8%)	26(35.6%)
Primary Eligibility *n*(%)								
None	0(0%)	3(30%)	0(0%)	3(9.7%)	0(0%)	3(7.9%)	13(100%)	3(4.1%)
SLD	74(97.4%)	0(0%)	52(94.5%)	22(71%)	45(93.8%)	29(76.3%)	0(0%)	61(83.6%)
OHI	0(0%)	5(50%)	1(1.8%)	4(12.9%)	1(2.1%)	4(10.5%)	0(0%)	5(6.8%)
ED	1(1.3%)	1(10%)	2(3.6%)	0(0%)	2(4.2%)	0(0%)	0(0%)	2(2.7%)
HI	1(1.3%)	1(10%)	0(0%)	2(6.5%)	0(0%)	2(5.3%)	0(0%)	2(2.7%)
Secondary Eligibility *n*(%)								
None	72(94.7%)	8(80%)	50(90.9%)	28(90.3%)	44(91.7%)	34(89.5%)	12(92.3%)	66(90.4%)
SLD	2(2.6%)	0(0%)	3(5.5%)	1(3.2%)	2(4.2%)	2(5.3%)	1(7.7%)	3(4.1%)
OHI	2(2.6%)	2(20%)	2(3.6%)	2(6.4%)	2(4.2%)	2(5.2%)	0(0%)	4(5.4%)

Note. Ell = English language learners; MET = multidisciplinary evaluation team; SLD = Specific Learning Disability; No SLD = no Specific Learning Disability; *n* = number of participants; *M* = mean; *SD* = standard deviation; None = no educational disability; OHI = Other Health Impairment; ED = Emotional Disturbance; HI = Hearing Impairment.

Instruments

Wechsler Intelligence Scales for Children-Fourth Edition. The WISC-IV is an individually administered standardized and norm-referenced IQ test composed of a standard battery of 10 core subtests (M = 10; SD = 3) that create four index composite scores and a Full Scale IQ Score (M = 100; SD = 15). The Verbal Comprehension Index (VCI) measures "verbal concept formation" (p. 4) and it is created by the Similarities, Vocabulary, and Comprehension subtests. The Perceptual Reasoning Index (PRI) measures "non-verbal and fluid reasoning" (p. 4) and is formed from the Block Design, Picture Concepts, and Matrix Reasoning subtests. The Working Memory Index (WMI) measures "working memory" (p. 4) and is created from the Digit Span and Coding subtests. Finally, the Processing Speed Index (PSI) measures "processing speed" (p. 4) and is formed from the Letter-Number Sequencing and Symbol Search subtests (Wechsler, 2003b).

The WISC-IV standardization sample includes 2,200 children ages 6 years and 0 months to 16 years and 11 months who represented the 2000 United States census stratified on age, sex, race, ethnicity, parent education level, and geographic region. Students "not fluent in English" (p. 24) were specifically excluded from participation in the WISC-IV standardization process (Wechsler, 2003b). WISC-IV scores exhibit good psychometric properties (Chen & Zhu, 2008; Edwards & Paulin, 2007; Launey, Carroll, & Van Horn; Watkins, Wilson, Kotz, Carbone, Babula; Wechsler, 2003). WISC-IV test developers gathered internal consistency and test-retest reliability estimates to support the reliability of the WISC-IV scores. The average internal consistency coefficients were 0.97 for the FSIQ, 0.94 for the VCI, 0.92 for the WMI, and 0.88 for the PSI and the median internal consistency coefficients for individual subtests ranged from 0.79 for Symbol Search and Cancellation to 0.90 for Letter-Number Sequencing. Additionally, test-retest reliability yielded a test-retest stability coefficient of 0.89 for the FSIQ, 0.89 for the VCI, 0.85 for PRI, 0.85 for WMI, and 0.79 for PSI on a sample of 243 children who were administered the WISC-IV twice at intervals ranging from 13 to 63 days. Validity evidence was gathered from a series of exploratory and confirmatory factor analyses. Results from an exploratory factor analysis identified factor loadings from the standard battery core subtests that matched the predicted factor structure of VCI, PRI, WMI, and PSI. In addition, results from a confirmatory factor analysis supported this structure (Wechsler, 2003b).

Wechsler Individual Achievement Tests-Third Edition. The Wechsler Individual Achievement Tests-Third Edition (WIAT-III) are a set of standardized achievement tests that can be administered individually or in small groups to students in preschool through twelfth grade or ages 4 years 0 month through 19 years 11 months. The WIAT-III consists of 16 subtests (M = 100; SD = 15) that comprise 8 composite scores (M = 100; SD = 15).

The WIAT-III standardization group consists of a sample of 2,775 students in preschool through twelfth grades representative of the 2005 U. S. census data stratified on grade, age, sex, race/ethnicity, parent education level, and geographic region. An equal number of female and male participants are included in the standardization sample. The WIAT-III exhibits good psychometric properties. Average internal-consistency reliability coefficients for all grade levels ranges from 0.69 to 0.97 for all subtest scores and from 0.91 to 0.98 for all composite scores in the Fall. In the Spring, average internal-consistency reliability coefficients for all grade levels ranges from 0.69 to 0.97 for all subtest scores and from 0.91 to 0.98 for all composite scores. Internal-consistency reliability coefficients were also provided by age group. Average internal-consistency reliability coefficients for all ages ranges from 0.69 to 0.97 for all subtest scores and from 0.91 to 0.98 for all composite scores. Additionally, test-retest stability coefficients ranged from 0.62 to 0.91 for all subtest scores and from 0.83 to 0.95 for composite scores for students in preschool through fifth grade at test-retest intervals that ranged from 2 to 32 days and from 0.69 to 0.94 for all subtest scores and from 0.87 to 0.96 for composite scores for students in sixth through twelfth grade at the same test-retest time intervals.

Validity evidence was gathered from an examination of the intercorrelations between the item scores, subtest scores, and the composite scores as well as from an examination of the relationship between scores from the WIAT-III and scores from other tests that measure achievement and ability. Intercorrelations ranged from 0.46 to 0.93 on the composite scores of the WIAT-III. Additionally, correlations between the WIAT-III and the WIAT-II ranged from 0.62 to 0.86 for common subtests and the corrected correlations range from 0.76 to 0.93 for composite scores. Finally, there is a positive correlation between the WIAT-III Total Achievement composite and the general intelligence score of the following standardized IQ tests: (a) the WISC-IV (r = 0.82), (b) the

Wechsler Adult Intelligence Scales-Fourth Edition (r = 0.82), (c) the Wechsler Nonverbal Scale (r = 0.60), and (d) the Differential Ability Scales-Second Edition (r = 0.67) (Breaux, 2009).

Woodcock-Johnson Tests of Achievement-Third Edition. The Woodcock-Johnson Tests of Achievement-Third Edition (WJ-III Ach) are a set of standardized achievement tests that can be administered individually or in small groups to students aged 2 years through 90+. The WJ-III Ach consists of 22 subtests (M = 100; SD = 15) that comprise 19 possible composite scores (M = 100; SD = 15).

The WJ-III Ach standardization group consists of a sample of 8,818 students in preschool through twelfth grades representative of the 2000 U. S. census data stratified on census region, community size, sex, race, Hispanic origin, type of school, type of college/university, education of adults, occupational status of adults, and occupation of adults in the labor force. An equal number of female and male participants are included in the standardization sample. The WJ-III Ach also demonstrates adequate psychometric properties.

The authors of the WJ-III Ach technical manual only published reliability and validity evidence on a selected number of subtests and composites due to the large number of subtest and composite scores available on the WJ-III Ach. Test-retest reliabilities over a time interval of 1 day for selected tests ranged from 0.76 to 0.95 for individuals aged 7 through 11, from 0.73 to 0.89 for individuals aged 14 through 17, and from 0.69 to 0.96 for individuals aged 26 through 79. Test-retest reliabilities ranged from 0.93 to 0.99 for selected composite scores for all age ranges over a time interval of 1 year. Inter-rater reliability ranged from 0.88 to 0.92 for all grade levels on the Writing Samples subtest and from 0.96 to 0.99 for all grade levels on the Writing Fluency subtest (McGrew & Woodcock, 2001).

Validity evidence was also gathered for the WJ-III Ach. Convergent validity was investigated through an examination of correlations between subtests from the WJ-III and subtests from the Kaufman Test of Educational Achievement (KTEA) and subtests from the Wechsler Individual Achievement Test (WIAT). Correlations between scores on the WJ-III Ach and the KTEA were moderate to high on tests that purportedly measure similar concepts. For example, correlations between the two achievement tests on reading related composite scores ranged from 0.44 to 0.81 for

all age ranges and correlations between the two achievement tests on math related composite scores ranged from 0.29 to 0.67. Correlations between scores on the WJ-III Ach and the WIAT for subtests that measure similar constructs were slightly higher. For example, correlations between reading related composite scores ranged from 0.67 to 0.82 and correlations between math related composite scores ranged from 0.59 to 0.70. In addition, results from an exploratory factor analysis determined that the factor loadings of the WJ-III Ach subtests matched the broad Cattell-Horn-Carroll abilities (McGrew & Woodcock, 2001).

In 2007, the normative comparison group was updated on the WJ-III Ach and additional reliability and validity evidence was gathered. The WJ-III Ach Normative Update participants consist of 8,782 students in preschool through college representative of the 2005 U. S. census data stratified on census region, community size, sex, race, Hispanic origin, type of school, type of college/university, education of adults, occupational status of adults, occupation of adults in the labor force, and foreign versus native born status (McGrew, Schrank, & Woodcock, 2007).

Similarly, only a selected number of reliability and validity statistics were published due to the large number of subtest and composite scores available on the WJ-III Ach Normative Update. Test-retest reliabilities over a time interval of 1 day for selected tests ranged from 0.75 to 0.94 for individuals aged 7 through 11, from 0.72 to 0.97 for individuals aged 14 through 17, and from 0.67 to 0.95 for individuals aged 26 through 79. Test-retest reliabilities ranged from 0.69 to 0.99 for selected composite scores for all age ranges over a time interval of 1 year. Inter-rater reliability ranged from 0.88 to 0.92 for all grade levels on the Writing Samples subtest and from 0.96 to 0.99 for all grade levels on the Writing Fluency subtest (McGrew et al., 2007).

Validity information for the WJ-III Ach Normative Update was gathered from: (a) a discussion of how the scores from the WJ-III Ach Normative Update subtests align with the Cattell-Horn-Carrol theory of intelligence and (b) a review of related research in the field of cognitive psychology (McGrew et al., 2007).

Procedure

It would be beyond the scope of this study to examine the validity of every cognitive test classified within a C-LIM. Therefore, the diagnostic utility of the WISC-IV C-LIM was investigated

because of the common use of the WISC-IV in applied settings (Kaufman & Lichtenberger, 2000) and its value as a diagnostic assessment instrument for children (Weiss, Beal, Saklofske, Alloway, & Prifitera, 2008). An illustration of the WISC-IV C-LIM is provided in figure 4.

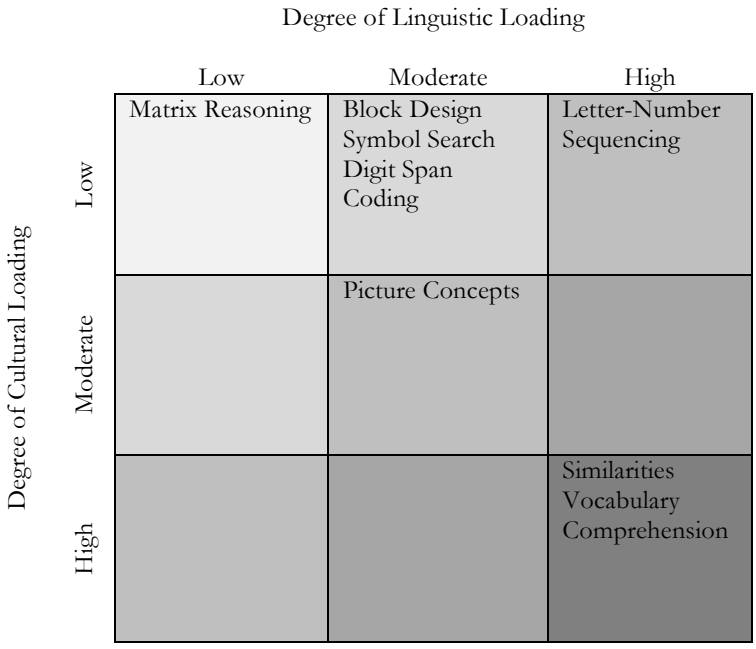

Figure 4. A C-LIM containing the WISC-IV standard battery according to Flanagan et al.'s (2007) hypothesized C-LTC system.

Participant data were collected systematically upon receipt of Arizona State University Institutional Review Board approval and the approval of the two school districts from which the data were collected. All of the active special education files for each school district were examined for the presence of WISC-IV scores and information was only collected from files that contained these scores. Data that were collected included demographic information, WISC-IV scores, standardized achievement scores, and the disability eligibility criteria of each individual participant.

Flanagan et al. (2007) claimed that scores from students who meet criteria for a Speech and Language Impairment, Autism, or an Intellectual Disability may exhibit a declining pattern of scores similar to the different profile hypothesized to exist amongst scores from culturally and linguistically diverse students without SLD. Therefore, students were included in the referred sample if they met the following criteria: (a) a parent home language survey indicated that Spanish was the primary language spoken at home, (b) they were not identified as students with Speech and Language

Impairments, Autism, or an Intellectual Disability by a school multidisciplinary evaluation team, (c) their file contained all ten core WISC-IV subtest scores, and (d) their file contained achievement subtests that correspond to IDEA (2004) SLD eligibility in the areas of basic reading skills, reading fluency, reading comprehension, mathematics calculation, mathematics problem solving, or written expression. Participants were excluded from the referred sample and from the WISC-IV normative sample if their WISC-IV FSIQ score was less than or equal to 73 to further ensure that participants did not include students who met criteria for an Intellectual Disability. This value was chosen because the WISC-IV FSIQ score contains a standard error of measurement that ranges from 2.68 to 3 points across all age ranges (Wechsler, 2003).

Next, scores from the referred sample of Ells were examined for the presence of SLD. Participants were identified as having SLD or not having SLD according to four different criteria currently used in applied practice in order to determine the degree to which varying SLD criteria changes the accuracy of the C-LIM decisions: (a) school multidisciplinary evaluation team decisions (IDEA, 2004), (b) the presence of an ability-achievement discrepancy using the regression formula (Mercer et al., 1996; Reschly & Hosp, 2004) and an ability-achievement regression of 0.60 (LD Technical Supplement, N. D.), (c) the presence of an ability-achievement discrepancy of one standard deviation (Mercer et al., 1996), and (d) the presence of an ability-achievement discrepancy of two standard deviations (Mercer et al., 1996). Treating all Ells as having SLD and treating all Ells as not having SLD formed two additional groups for the sensitivity analysis. Finally, participants were labeled different or disordered according to the criteria outlined by Flanagan et al. (2007) for the WISC-IV C-LIM.

Analyses

Means and standard deviations for all WISC-IV subtest, composite, and FSIQ scores were calculated first. Additionally, the degree to which WISC-IV subtest, composite, and FSIQ scores differed significantly between each of the participant subgroups was analyzed using a one-way ANOVA. The Levene's Test of Homogeneity of Variances was examined to test the degree to which the score variances between the participant subgroups differed significantly. A one-way ANOVA was also conducted to test if mean subtest, index, and FSIQ scores varied significantly between participant

subgroups using the Welch approximate *F* test to determine overall significance and the Dunnett's C test to evaluate post-hoc comparisons because those tests do not assume homogeneity of variances. In addition, the Bonferonni correction was used to maintain an experimentwise error rate of .05 in order to reduce type I error from conducting multiple statistical tests. The resultant alpha level for each individual test was set at .003.

Frequencies of the WISC-IV C-LIM decisions (i.e., different and disordered) were compared to the true states of cultural and linguistic diversity from the WISC-IV normative sample for each of the following pairs of participants in order to compute true positive and false positive rates for every possible cut-score to plot on a ROC graph: (a) Ells with SLD and the WISC-IV normative sample, (b) Ells without SLD and the WISC-IV normative sample, and (c) Ells with SLD and Ells without SLD. It is important to note that Ells were not included in the WISC-IV standardization process and are therefore culturally and linguistically different from the WISC-IV normative sample (2003b). However, Flanagan et al. (2007) stated that an absence of the declining pattern of scores for culturally and linguistically diverse students indicates that, "the most likely reason will be due to learning disability because the results will vary more as a function of which test scores happen to be low and what constructs those tests are designed to measure," (p. 197) which infers that Flanagan et al. hypothesize that scores from students who are diagnosed with SLD should follow the disordered pattern. C-LIM decisions were compared between samples of Ells with SLD and the WISC-IV normative sample in order to test that statement. Thus, the true state of Ells with SLD was treated as different and the true state of the WISC-IV normative sample was treated as disordered for those comparisons. The comparisons between the C-LIM decisions for the samples of Ells with SLD and the samples of Ells without SLD were conducted for a similar purpose, but with a modified protocol. For those comparisons, the true state of Ells with SLD was treated as disordered and the true state of Ells without SLD was treated as different. Lastly, the comparison between the samples of Ells without SLD and the WISC-IV normative sample was conducted to test Flanagan et al.'s primary claim that the C-LIM assists practitioners in answering the question of "difference versus disorder" (p. 175). As stated previously, Ells were not included in the WISC-IV standardization process (Wechsler, 2003b), therefore the true state of Ells without SLD is different and the true state of the

WISC-IV normative sample is disordered. Table 2 contains a list of the WISC-IV C-LIM predicted score patterns and true states of cultural and linguistic diversity from the WISC-IV normative sample for each comparison.

Table 2

WISC-IV C-LIM Predicted Score Patterns for 86 Ells Tested for Special Education Eligibility and the WISC-IV Normative Sample

	C-LIM decisions		True states	
Participant subgroup	Different	Disordered	Different	Disordered
Ells with SLD vs. WISC-IV normative sample				
Ells with SLD		X	X	
WISC-IV normative sample		X		X
Ells without SLD vs. WISC-IV normative sample				
Ells without SLD	X		X	
WISC-IV normative sample		X		X
Ells with SLD vs. Ells without SLD				
Ells with SLD		X		X_a
Ells without SLD	X		X	

Note. C-LIM = Culture-Language Interpretive Matrix; Ells = English language learners; SLD = Specific Learning Disabilities; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition.
[a]The true state of Ells with SLD is treated as disordered for this comparison in order to test Flanagan et al.'s (2007) claim that when the different pattern does not emerge in the scores of students who are culturally and linguistically diverse, "the most likely reason will be due to learning disability" (p. 197).

Flanagan et al. (2007) did not provide a specific numerical value for the minimum discrepancy required between cells for a score pattern to be designated different, rather they stated:

> What is important to note is not the normative values of the scores (i.e., average, low average), but rather the relationships between the scores and the degree to which they form a pattern than is either consistent or inconsistent with the pattern of performance predicted by the matrix (Flanagan et al., 2007, p. 181).

Therefore, a continuous outcome variable was created to test the ability of the C-LIM to accurately distinguish between groups of students who are different and disordered across every possible cut score that emerged between the cells down the diagonal of the matrix. The minimum discrepancy that emerged between cells down the diagonal of the matrix ranged from 0 to 26 points. Then, the AUC was used to interpret the ROC. An AUC value is interpreted as the probability that diagnostic test results from a randomly selected individual who is positive for a trait (i.e., different) will yield a higher value than a randomly selected individual who does not have the trait (i.e., disordered) across every possible cut-score (Metz, 1978; Pepe, 2003; Streiner & Cairney, 2007; Swets, 1988). Swets (1988) and Streiner and Cairney (2007) suggested that AUC values between .50 and .70 characterize

low accuracy, AUC values between .70 and .90 characterize medium accuracy, and AUC values between .90 and 1.00 characterize high accuracy.

Both nonparametric (Bamber, 1975; Hanley & McNeil, 1982) and parametric (Dorfman & Alf, 1969; Metz, 1978) methods can be used to compute the AUC. Nonparametric approaches do not rely on distributional assumptions and can be used with small sample sizes (Hanley & McNeil, 1982). Parametric approaches are based on normal distributional assumptions and produce a smooth ROC curve. However, both approaches yield similar results and differences between the parameters obtained from the two approaches are often negligible. This is especially true when the number of cut-scores increase and span a larger portion of the ROC space (Centor & Schwartz, 1985). The nonparametric approach was applied using SPSS version 20 software because this approach is more appropriate for use with smaller samples and it does not require adherence to strict distributional assumptions (Bamber, 1975; Hanley & McNeil, 1982).

A binary AUC calculation was also computed for: (a) each comparison group (i.e., different or disordered) and (b) each degree of discrepancy that emerged between cells down the diagonal of the matrix rounded to the lowest whole number (i.e., AUC when the discrepancy ≥ 1 point, AUC when the discrepancy ≥ 2 points, etc). This analyses was conducted to directly evaluate the discriminative ability of the C-LIM as it is intended to be used (i.e., single "yes/no" cut-score according to the presence of the declining pattern) and to evaluate the discriminative ability of the C-LIM based on the magnitude of the discrepancy that emerged between the cells down the diagonal of the matrix (i.e., AUC value when the discrepancy between cells down the diagonal ≥ 1 point, AUC value when the discrepancy between cells down the diagonal ≥ 2 points, etc…). The AUC value of a single cut-score can be approximated by dividing the region under the ROC curve into two right triangles and a rectangle: (a) the first right triangle has sides with lengths of 1 – specificity and sensitivity, (b) the second right triangle has sides with lengths of specificity and 1 – sensitivity, and (c) the rectangle has sides of length specificity and sensitivity. Next, the area of each geometric shape is computed and summed. The sum of the three areas represents the approximate AUC value for a single cut-score (Cantor & Kattan, 2000).

A power analysis was conducted to determine the appropriate sample sizes necessary to achieve a medium effect size, $d = 0.30$. This value was chosen because effect sizes for IQ subtest analysis range from small to medium. Cohen (1992) suggested the following effect size values for chi-square goodness of fit tests and contingency tables when the power of a test is 0.80: (a) $d = 0.10$ for a small effect, (b) $d = 0.30$ for a medium effect, and (c) $d = 0.50$ for a large effect. When the power of a test is 0.80, the contingency table has $df = 1$, and alpha is set at 0.05, 785 participants would be needed to detect a small effect and 87 participants would be needed to detect a medium effect (Cohen, 1992). Cohen's *d* is an effect size appropriate for use with comparisons between the scores from two normally distributed samples; however, the AUC is a measure of the probability that a score drawn at random from a truly clinical group is higher than that drawn from a truly non-clinical group (Metz, 1978; Swets, 1988; Swets, Dawes, & Monahan, 2000). According to Rice and Harris (2005), a researcher has the power to detect an AUC of 0.59 or higher when $d = 0.31$. Unfortunately, this number was not reached.

Chapter 3

Results

Descriptive Statistics

WISC-IV subtest, index, and FSIQ scores from the Ells were slightly lower and varied more than scores from the WISC-IV normative sample. However, it is not uncommon for referred samples to share these characteristics (Canivez & Watkins, 1998; Watkins, 2010). Results of the one-way ANOVA indicated that average subtest, index, and FSIQ scores were significantly different between participant subgroups on all subtests, indices, and FSIQ scores: Block Design $F(8, 85.31) = 15.33$, $p < .001$, Similarities $F(8, 85.62) = 55.29$, $p < .001$, Digit Span $F(8, 85.69) = 64.79$, $p < .001$, Picture Concepts $F(8, 84.92) = 3.46$, $p = .002$, Coding $F(8, 84.59) = 11.35$, $p < .001$, Vocabulary $F(8, 85.59) = 91.32$, $p < .001$, Letter-Number Sequencing $F(8, 84.89) = 35.48$, $p < .001$, Matrix Reasoning $F(8, 85.39) = 9.73$, $p < .001$, Comprehension $F(8, 85.05) = 26.23$, $p < .001$, Symbol Search $F(8, 84.71) = 11.23$, $p < .001$, Verbal Comprehension Index $F(8, 85.83) = 81.30$, $p < .001$, Perceptual Reasoning Index $F(8, 85.98) = 17.74$, $p < .001$, Working Memory Index $F(8, 85.23) = 66.23$, $p < .001$, Processing Speed Index $F(8, 84.79) = 15.45$, $p < .001$, Full Scale Intelligence Quotient $F(8, 86.51) = 107.19$, $p < .001$.

Post-hoc tests conducted with the Dunnet's C test indicated that the samples of Ells disaggregated by SLD criteria were significantly different from the WISC-IV normative sample, but not significantly different from each other. Table 3 provides detailed results.

Table 3

Means and Standard Deviations of the WISC-IV Subtest, Index, and FSIQ Scores for 86 Ells Tested for Special Education Eligibility Disaggregated by SLD Criteria and the WISC-IV Normative Sample

	WISC-IV	MET decision		Regression discrepancy		One standard deviation discrepancy		Two standard deviation discrepancy	
	No SLD	SLD	No SLD	SLD	No SLD	SLD	No SLD	SLD	No SLD
	n = 2,033	*n* = 76	*n* = 10	*n* = 55	*n* = 31	*n* = 48	*n* = 38	*n* = 13	*n* = 73
Test	*M*(*SD*)	*M*(*SD*)	*M*(*SD*)	*M*(*SD*)	*M*(*SD*)	*M*(*SD*)	*M*(*SD*)	*M*(*SD*)	*M*(*SD*)
BD	10.3(2.8)	9.0(2.7)	7.6(2.0)	9.2(2.6)	8.0(2.7)*	9.5(2.7)	7.9(2.5)*	11.3(2.1)	8.4(2.6)*
SI	10.3(2.8)	7.4(2.3)*	7.9(2.0)	7.5(2.5)*	7.5(1.8)*	7.9(2.5)*	6.9(1.9)*	8.9(3.0)	7.2(2.1)*
DS	10.3(2.8)	7.5(2.0)*	6.0(2.7)	7.3(2.2)*	7.4(1.9)*	7.4(2.2)*	7.2(2.0)*	8.5(2.4)	7.1(2.0)*
PCn	10.3(2.8)	9.6(2.6)	8.9(3.3)	9.7(2.9)	9.2(2.4)	9.7(3.0)	9.3(2.3)	10.5(2.0)	9.3(2.8)
CD	10.3(2.8)	8.7(2.7)*	9.6(3.8)	9.3(2.9)	8.0(2.6)*	9.5(3.0)	8.0(2.5)*	9.2(3.0)	8.8(2.8)*
VC	10.3(2.7)	6.8(2.1)*	7.3(2.3)	7.0(2.3)*	6.7(1.8)*	7.2(2.3)*	6.5(1.8)*	8.1(2.8)	6.7(1.9)*
LN	10.3(2.8)	7.8(2.5)*	6.5(3.8)	7.7(2.7)*	7.5(2.8)*	8.1(2.4)*	7.0(3.0)*	9.6(2.0)	7.3(2.7)*
MR	10.3(2.8)	9.2(2.4)	8.1(2.4)	9.4(2.3)	8.5(2.5)	9.5(2.2)	8.5(2.5)	10.1(2.0)	8.9(2.4)*
CO	10.4(2.7)	8.4(2.1)*	8.3(3.5)	8.5(2.3)*	8.2(2.4)*	8.8(2.2)*	7.9(2.3)*	10.1(2.7)	8.1(2.1)*
SS	10.3(2.7)	8.9(2.7)*	9.4(2.9)	9.1(3.0)	8.7(2.3)	9.1(3.0)	8.7(2.4)	11.1(3.5)	8.6(2.4)*
VCI	101.2(13.4)	85.9(10.3)*	87.8(9.8)	86.8(11.1)*	85.0(8.6)*	88.6(10.9)*	83.0(8.5)*	94.9(14.2)	84.6(8.6)*
PRI	101.8(13.3)	95.8(11.1)*	89.0(9.5)	97.2(11.4)	91.2(9.7)*	98.0(11.5)	91.2(9.5)*	103.7(7.5)	93.5(11.0)*
WMI	101.1(13.4)	86.3(10.7)	78.0(14.4)*	85.6(12.1)*	84.8(10.4)*	87.0(11.4)*	83.2(11.3)*	93.8(10.8)	83.8(11.0)*
PSI	101.8(13.5)	93.6(13.1)*	97.3(13.4)	95.9(13.8)	90.7(11.4)*	96.6(13.9)	90.8(11.5)*	100.5(17.3)	92.9(12.0)*
FSIQ	102.2(13.0)	87.6(9.3)*	84.6(9.5)	88.9(10.0)*	84.4(7.0)*	90.5(9.9)*	83.2(6.6)*	98.4(10.1)	85.3(7.6)*

Note. MET = Multidisciplinary Evaluation Team; SLD = Specific Learning Disability; No LD = No Specific Learning Disability; *n* = Number of participants; *M* = Mean; *SD* = Standard Deviation; BD = Block Design; SI = Similarities; DS = Digit Span; PCn = Picture Concepts; CD = Coding; VC = Vocabulary; LN = Letter-Number Sequencing; MR = Matrix Reasoning; CO = Comprehension; SS = Similarities; VCI = Verbal Comprehension Index; PRI = Perceptual Reasoning Index; PSI = Processing Speed Index; WMI = Working Memory Index; FSIQ = Full Scale Intelligence Quotient.

*$p < .003$

Flanagan et al.'s (2007) disordered profile (i.e., no pattern of decline) emerged in the mean WISC-IV subtest scores from the WISC-IV normative sample. However, Flanagan et al.'s disordered profile also emerged in the mean WISC-IV subtest scores from all of the samples containing Ells (i.e., Ells with SLD, Ells without SLD, and all 86 Ells grouped together). This indicates that, on average, the different profile did not exist in the scores for these students. Likewise, the individual decisions based on C-LIM criteria produced a high frequency of true negative decisions (i.e., number of students diagnosed as disordered who came from the WISC-IV normative sample) and a low frequency of true positive decisions (i.e., number of students diagnosed as different who came from the samples of Ells) when comparing Ells with and without SLD to the WISC-IV normative sample. The C-LIM produced a high frequency of false negative decisions as a result, which indicates that it frequently diagnosed students as disordered who came from the samples of Ells regardless of the presence of SLD or the criteria used to identify it. These trends remained in place for the comparisons between Ells with SLD and Ells without SLD. Figures 5-19 illustrate contingency tables containing the frequencies of true positive, true negative, false positive, and false negative decisions for each of the participant subgroups contained in each paired comparison.

Figure 5

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 76 Ells Identified as Having SLD by a MET Decision and the WISC-IV Normative Sample

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells with SLD	7_a	69_d	76
	WISC-IV normative sample	100_c	$1{,}933_b$	2,033
	Total	107	2,002	2,109

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability; MET = multidisciplinary evaluation team.
[a]These frequencies represent true positive decisions.
[b]These frequencies represent true negative decisions.
[c]These frequencies represent false positive decisions.
[d]These frequencies represent false negative decisions.

Figure 6

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 55 Ells Identified as Having SLD by an Ability-Achievement Discrepancy Using the Regression Equation and the WISC-IV Normative Sample

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells with SLD	5_a	50_d	55
	WISC-IV normative sample	100_c	$1{,}933_b$	2,033
	Total	105	1,983	

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
[a]These frequencies represent true positive decisions.
[b]These frequencies represent true negative decisions.
[c]These frequencies represent false positive decisions.
[d]These frequencies represent false negative decisions.

Figure 7

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 48 Ells Identified as Having SLD by an Ability-Achievement Discrepancy of One Standard Deviation and the WISC-IV Normative Sample

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells with SLD	5_a	43_d	48
	WISC-IV normative sample	100_c	$1{,}933_b$	2,033
	Total	105	1,976	2,081

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
[a]These frequencies represent true positive decisions.
[b]These frequencies represent true negative decisions.
[c]These frequencies represent false positive decisions.
[d]These frequencies represent false negative decisions.

Figure 8

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 13 Ells Identified as Having SLD by an Ability-Achievement Discrepancy of Two Standard Deviations and the WISC-IV Normative Sample

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells with SLD	1_a	12_d	13
	WISC-IV normative sample	100_c	$1{,}933_b$	2,033
	Total	101	1,945	2,046

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
$_a$These frequencies represent true positive decisions.
$_b$These frequencies represent true negative decisions.
$_c$These frequencies represent false positive decisions.
$_d$These frequencies represent false negative decisions.

Figure 9

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 86 Ells Tested for Special Education Eligibility Treated as SLD and the WISC-IV Normative Sample

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells with SLD	9_a	77_d	86
	WISC-IV normative sample	100_c	$1{,}933_b$	2,033
	Total	109	2,010	2,119

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
$_a$These frequencies represent true positive decisions.
$_b$These frequencies represent true negative decisions.
$_c$These frequencies represent false positive decisions.
$_d$These frequencies represent false negative decisions.

Figure 10

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 10 Ells Identified as Not Having SLD by a MET Decision and the WISC-IV Normative Sample

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells without SLD	0_a	10_d	10
	WISC-IV normative sample	100_c	$1{,}933_b$	2,033
	Total	100	1,943	2,043

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
$_a$These frequencies represent true positive decisions.
$_b$These frequencies represent true negative decisions.
$_c$These frequencies represent false positive decisions.
$_d$These frequencies represent false negative decisions.

Figure 11

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 31 Ells Identified as Not Having SLD by an Ability-Achievement Discrepancy Using the Regression Equation and the WISC-IV Normative Sample

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells without SLD	2_a	29_d	31
	WISC-IV normative sample	100_c	$1{,}933_b$	2,033
	Total	102	1,962	2,064

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
$_a$These frequencies represent true positive decisions.
$_b$These frequencies represent true negative decisions.
$_c$These frequencies represent false positive decisions.
$_d$These frequencies represent false negative decisions.

Figure 12

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 38 Ells Identified as Not Having SLD by an Ability-Achievement Discrepancy of One Standard Deviation and the WISC-IV Normative Sample

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells without SLD	2_a	36_d	38
	WISC-IV normative sample	100_c	$1{,}933_b$	2,033
	Total	102	1.969	2,071

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
$_a$These frequencies represent true positive decisions.
$_b$These frequencies represent true negative decisions.
$_c$These frequencies represent false positive decisions.
$_d$These frequencies represent false negative decisions.

Figure 13

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 73 Ells Identified as Not Having SLD by an Ability-Achievement Discrepancy of Two Standard Deviations and the WISC-IV Normative Sample

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells without SLD	6_a	67_d	73
	WISC-IV normative sample	100_c	$1{,}933_b$	2,033
	Total	106	2,000	2,106

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
$_a$These frequencies represent true positive decisions.
$_b$These frequencies represent true negative decisions.
$_c$These frequencies represent false positive decisions.
$_d$These frequencies represent false negative decisions.

Figure 14

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 86 Ells Tested for Special Education Eligibility Treated as Not Having SLD and the WISC-IV Normative Sample

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells without SLD	9_a	77_d	86
	WISC-IV normative sample	100_c	$1{,}933_b$	2,033
	Total	109	2,010	2,119

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
$_a$These frequencies represent true positive decisions.
$_b$These frequencies represent true negative decisions.
$_c$These frequencies represent false positive decisions.
$_d$These frequencies represent false negative decisions.

Figure 15

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 76 Ells Identified as Having SLD and 10 Ells Identified as Not Having SLD by a MET Decision

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells without SLD	0_a	10_d	10
	Ells with SLD	7_c	69_b	76
	Total	7	79	86

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability; MET = multidisciplinary evaluation team.
$_a$These frequencies represent true positive decisions.
$_b$These frequencies represent true negative decisions.
$_c$These frequencies represent false positive decisions.
$_d$These frequencies represent false negative decisions.

Figure 16

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 31 Ells Identified as Having SLD and 55 Ells Identified as Not Having SLD by an Ability-Achievement Discrepancy Using the Regression Equation

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells without SLD	5_a	50_d	55
	Ells with SLD	2_c	29_b	31
	Total	7	79	86

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
$_a$These frequencies represent true positive decisions.
$_b$These frequencies represent true negative decisions.
$_c$These frequencies represent false positive decisions.
$_d$These frequencies represent false negative decisions.

Figure 17

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 48 Ells Identified as Having SLD and 38 Ells Identified as Not Having SLD by an Ability-Achievement Discrepancy of One Standard Deviation

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells without SLD	2_a	36_d	38
	Ells with SLD	5_c	43_b	48
	Total	7	79	86

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
$_a$These frequencies represent true positive decisions.
$_b$These frequencies represent true negative decisions.
$_c$These frequencies represent false positive decisions.
$_d$These frequencies represent false negative decisions.

Figure 18

Contingency Table Comparing the Frequencies of the C-LIM Decisions to the Frequencies of the True State of Cultural and Linguistic Diversity Status for 13 Ells Identified as Having SLD and 73 Ells Identified as Not Having SLD by an Ability-Achievement Discrepancy of Two Standard Deviations

		C-LIM Decisions		
		Different (*n*)	Disordered (*n*)	**Total**
Actual Traits	Ells without SLD	6_a	67_d	73
	Ells with SLD	1_c	12_b	13
	Total	7	79	86

Note. N = number of participants; % = percent of participants; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; Ell = English language learners; SLD = Specific Learning Disability.
[a]These frequencies represent true positive decisions.
[b]These frequencies represent true negative decisions.
[c]These frequencies represent false positive decisions.
[d]These frequencies represent false negative decisions.

Diagnostic Utility Statistics

ROC analyses were conducted to compare the frequencies of the WISC-IV C-LIM decisions (i.e., different and disordered) to the true states of cultural and linguistic diversity from the WISC-IV normative sample for each of the following pairs of participants: (a) Ells with SLD and the WISC-IV normative sample, (b) Ells without SLD and the WISC-IV normative sample, and (c) Ells with SLD and Ells without SLD. Flanagan et al. (2007) claimed that the C-LIM different profile emerges when students are culturally and linguistically different from the normative comparison group of a given IQ test. However, it was hypothesized that a comparison between a sample of Ells with SLD and the WISC-IV normative sample would yield AUC values between 0.50 and 0.70 in the low accuracy range (Streiner & Cairney, 2007; Swets, 1988) because the different profile has not been empirically supported by studies investigating the C-LIM score patterns for culturally and linguistically diverse students (Kranzler et al., 2010; Tychanska, 2009; Verderosa, 2007). Results indicated AUC values ranging from 0.51 to 0.53 for the comparisons between the samples of Ells with SLD and the WISC-IV normative sample regardless of the manner by which SLD was defined. This indicates that the C-LIM would accurately distinguish between a randomly selected participant from the referred sample

of Ells with SLD and a randomly selected participant from the simulated WISC-IV normative sample 51% to 53% of the time. Figures 19-23 illustrate the respective ROC curves for these analyses.

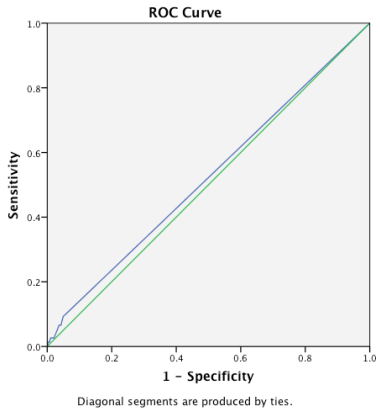

Figure 19. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as SLD by a multidisciplinary evaluation team (n = 76) and the WISC-IV normative sample (n = 2,033).

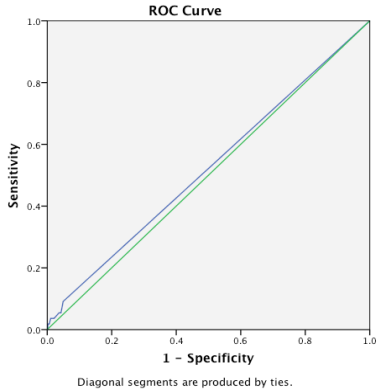

Figure 20. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as SLD by an ability-achievement discrepancy using the regression equation (n = 55) and the WISC-IV normative sample (n = 2,033).

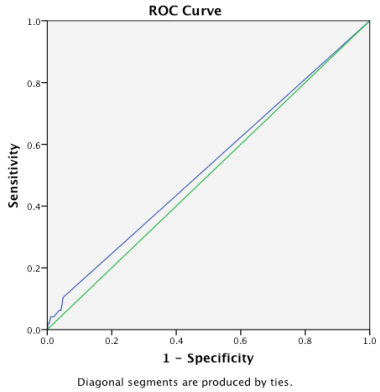

Figure 21. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as SLD by an ability-achievement discrepancy of 1 standard deviation (n = 48) and the WISC-IV normative sample (n = 2,033).

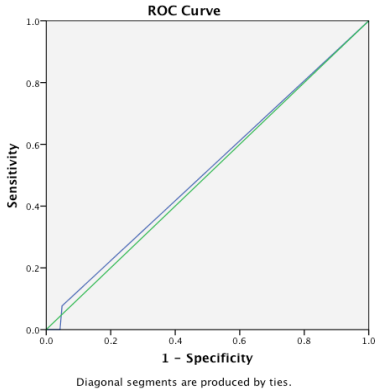

Figure 22. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as SLD by an ability-achievement discrepancy of 2 standard deviations (n = 13) and the WISC-IV normative sample (n = 2,033).

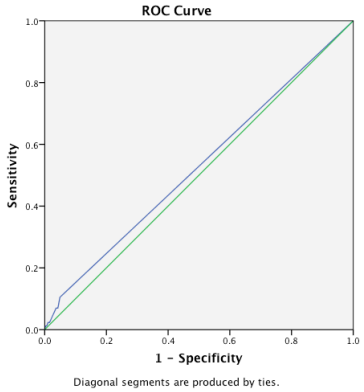

Figure 23. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells treated as SLD (n = 86) and the WISC-IV normative sample (n = 2,033).

Contrary to the assertions of Flanagan et al. (2007), a comparison between scores from a sample of Ells without SLD and scores from the WISC-IV normative sample were also hypothesized to yield AUC values between 0.50 and 0.70 in the low accuracy range .for the same rationale. Results of the ROC analyses supported this hypothesis. AUC values ranged from 0.48 to 0.53 regardless of the SLD criteria imposed. This indicates that the C-LIM would accurately distinguish between a randomly selected participant from the referred sample of English language learners without SLD and a randomly selected participant from the simulated WISC-IV normative sample 48 to 53% of the time, despite the criteria used to define SLD. Figures 24-28 illustrate the respective ROC curves for these analyses.

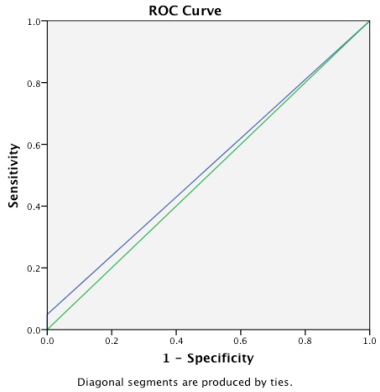

Figure 24. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as not having SLD by a multidisciplinary evaluation team decision (n = 10) and the WISC-IV normative sample (n = 2,033).

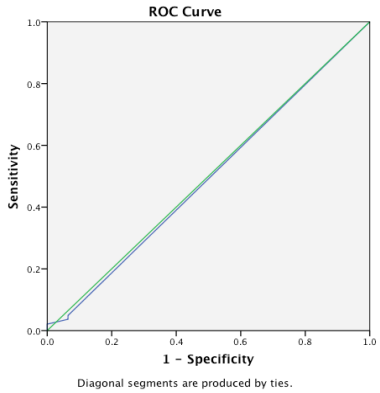

Figure 25. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as not having SLD by an ability-achievement discrepancy using the regression equation (n = 31) and the WISC-IV normative sample (n = 2,033).

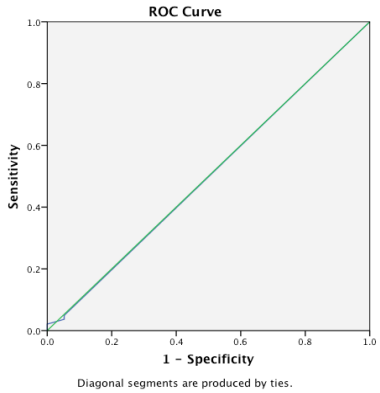

Figure 26. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as not having SLD by an ability-achievement discrepancy of 1 standard deviation (n = 38) and the WISC-IV normative sample (n = 2,033).

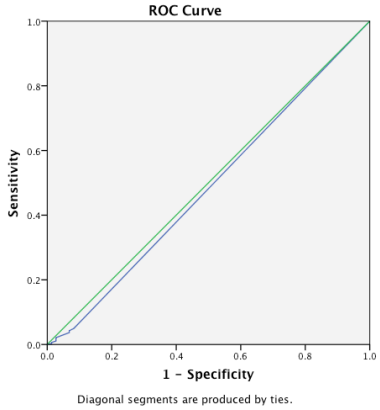

Figure 27. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as not having SLD by an ability-achievement discrepancy of 2 standard deviations (n = 73) and the WISC-IV normative sample (n = 2,033).

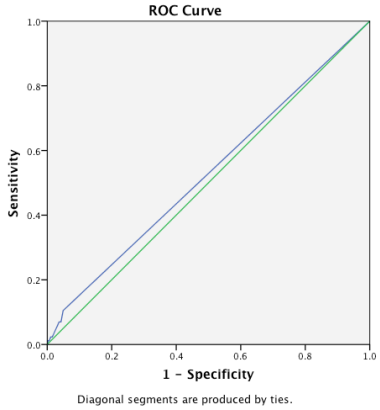

Figure 28. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells treated as not having SLD (n = 86) and the WISC-IV normative sample (n = 2,033).

The final comparison conducted was between the scores of Ells identified as SLD and the scores of Ells identified as not SLD. AUC values were hypothesized to fall between 0.50 and 0.70 in the low accuracy range (Streiner & Cairney, 2007; Swets, 1988) for similar reasons as those listed for the previous comparisons. Results indicated AUC values between 0.49 and 0.55. This means that the C-LIM would accurately distinguish between a randomly selected participant from the referred sample of Ells with SLD and a randomly selected participant from the referred sample of Ells without SLD 49 to 55% of the time, regardless of the manner in which SLD is defined. Figures 29-32 illustrate the respective ROC curves for these analyses.

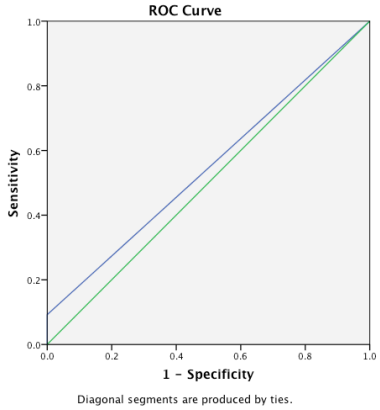

Figure 29. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as SLD (n = 76) and as not having SLD (n = 10) by a multidisciplinary evaluation team decision.

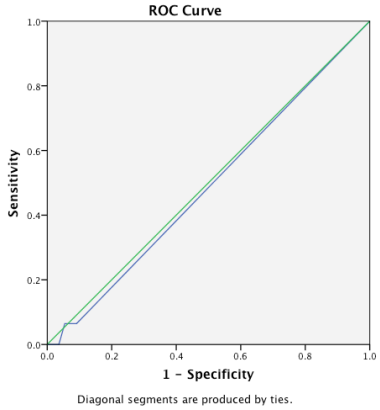

Figure 30. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as SLD (n = 55) and as not having SLD (n = 31) by an ability-achievement discrepancy using the regression equation.

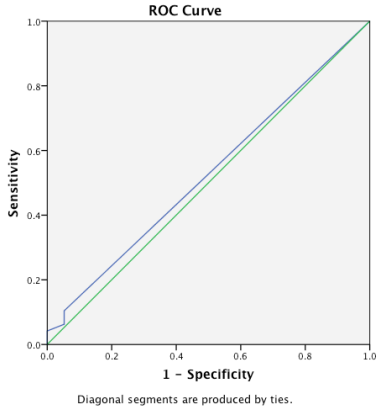

Figure 31. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as SLD (n = 48) and as not having SLD (n = 38) by an ability-achievement discrepancy of 1 standard deviation.

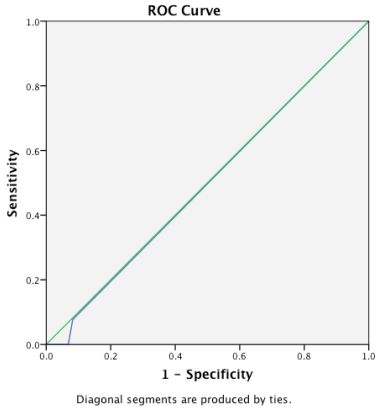

Figure 32. ROC curve comparing true-positive and false-positive rates from a referred sample of Ells identified as SLD (n = 13) and as not having SLD (n = 73) by an ability-achievement discrepancy of 2 standard deviations.

These ROC analyses were supplemented by a binary AUC calculation for: (a) each comparison group and (b) each degree of discrepancy that emerged between cells down the diagonal of the matrix. Binary ROC analyses for each comparison group produced AUC values that ranged between 0.45 and 0.53. Table 4 contains the AUC values for each paired comparison. Additionally, the AUC values for each degree of discrepancy that emerged between cells down the diagonal of the matrix are provided in table 5. AUC values ranged from 0.46 to 0.52 based on the magnitude of the discrepancy that emerged between cells down the diagonal of the matrix for all three main comparisons regardless of how SLD was defined. All values obtained from the binary ROC analyses characterize low diagnostic accuracy (Streiner & Cairney, 2007; Swets, 1988).

Table 4

Sensitivity, Specificity, and AUC Values for Binary ROC Analyses Conducted for Each Comparison Group Disaggregated by SLD Criteria

SLD criteria	Different[a] *n*	Disordered[b] *n*	Sensitivity	Specificity	AUC
Ells with SLD vs. WISC-IV normative sample					
MET decision	76	2,033	0.09	0.95	0.52
Regression discrepancy	55	2,033	0.09	0.95	0.52
1 SD discrepancy	48	2,033	0.10	0.95	0.53
2 SD discrepancy	13	2,033	0.08	0.95	0.51
All Ells	86	2,033	0.10	0.95	0.53
Ells without SLD vs. WISC-IV normative sample					
MET decision	10	2,033	<0.00	0.95	0.48
Regression discrepancy	31	2,033	0.06	0.95	0.51
1 SD discrepancy	38	2,033	0.05	0.95	0.50
2 SD discrepancy	73	2,033	0.08	0.95	0.52
All Ells	86	2,033	0.10	0.95	0.53
Ells with SLD vs. Ells without SLD					
MET decision	10	76	<0.00	0.91	0.45
Regression discrepancy	31	55	0.09	0.94	0.51
1 SD discrepancy	38	48	0.05	0.90	0.47
2 SD discrepancy	73	13	0.08	0.92	0.50

Note. AUC = Area Under the Curve; ROC = Receiver Operating Characteristic Curve; SLD = Specific Learning Disability; *n* = number of participants; MET = multidisciplinary evaluation team; Ell = English language learner; SD = standard deviation.

[a]This refers to the true state of participants in each subgroup who are considered to be culturally and linguistically different from the WISC-IV normative sample.

[b]This refers to the true state of participants in each subgroup who are considered to be culturally and linguistically similar to the WISC-IV normative sample.

Table 5

Binary AUC Values and Percent of Participants' Whose Scores Exhibited the C-LIM Different Profile Based on the Magnitude of the Discrepancy Between Cells Down the Diagonal of the Matrix

	WISC-IV	MET decision					Regression Discrepancy					One standard deviation discrepancy					Two standard deviation discrepancy					All Ells	
Cut-score	%	SLD %	No SLD %	AUC_a	AUC_b	AUC_c	SLD %	No SLD %	AUC_a	AUC_b	AUC_c	SLD %	No SLD %	AUC_a	AUC_b	AUC_c	SLD %	No SLD %	AUC_a	AUC_b	AUC_c	%	AUC_d
1	0.6	2.6	0	0.49	0.49	0.49	3.6	0	0.49	0.50	0.52	4.2	0	0.48	0.50	0.48	7.7	1.4	0.46	0.50	0.47	3.5	0.51
2	0	0	0				0	0				0	0				0	0				0	
3	0.6	0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0.50
4	0	0	0				0	0				0	0				0	0				0	
5	1.6	3.9	0	0.49	0.50	0.48	1.8	6.5	0.50	0.52	0.48	2.1	5.3	0.50	0.52	0.52	0	4.1	0.51	0.51	0.52	3.5	0.51
6	0.5	0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		1.2	0.50
7	0	0	0				0	0				0	0				0	0				0	
8	0.5	0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0.50
9	0	0	0				0	0				0	0				0	0				0	
10	0.4	1.3	0	0.50	0.49	0.49	1.8	0	0.49	0.50	0.51	2.1	0	0.49	0.50	0.49	0	1.4	0.50	0.50	0.51	1.2	0.50
11	0.3	0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0.50
12	0	0	0				0	0				0	0				0	0				0	
13	0.1	0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0.50
14	0	0	0				0	0				0	0				0	0				0	
15	0.1	1.3	0	0.49	0.49	0.49	1.8	0	0.49	0.50	0.51	2.1	0	0.49	0.50	0.49	0	1.4	0.50	0.51	0.51	1.2	0.51
16	0	0	0				0	0				0	0				0	0				0	
17	0	0	0				0	0				0	0				0	0				0	
18	<0.0	0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0.50
19	0	0	0				0	0				0	0				0	0				0	
20	<0.0	0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0.50
21	0	0	0				0	0				0	0				0	0				0	
22	0	0	0				0	0				0	0				0	0				0	
23	0	0	0				0	0				0	0				0	0				0	
24	0	0	0				0	0				0	0				0	0				0	
25	<0.0	0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0	0.50	0.50		0	0.50

Note. WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; MET = Multidisciplinary Evaluation Team; SLD = Specific Learning Disability; No SLD = No Specific Learning Disability; % = percent of individuals who exhibited a cut-score at the specified value; AUC = Area Under the Curve; Ell = English language learner.

[a]AUC values from a comparison between participants from the WISC-IV normative sample and Ells with SLD.

[b]AUC values from a comparison between participants from the WISC-IV normative sample and Ells without SLD.

[c]AUC values from a comparison between Ells with SLD and Ells without SLD.

[d]AUC values from a comparison between all Ells considered either SLD or not SLD and the WISC-IV normative sample.

Chapter 4

Discussion

The purpose of the present study was to determine the degree to which the WISC-IV C-LIM can accurately distinguish between scores from Ells who have SLD, scores from Ells who do not have SLD, and scores from the WISC-IV normative sample. Results from descriptive statistics indicated that significant mean differences did not emerge between groups of Ells with and without SLD on WISC-IV subtest, composite, or FSIQ scores. However, scores from the referred sample were significantly lower than scores from the WISC-IV normative sample. Flanagan et al. (2007) claimed that research on IQ test scores, "produced a rather strong consensus that bilinguals tended to perform about one standard deviation below the mean of monolinguals" (p. 169) and this argument was used to support their categorization system of subtests as having low, moderate, or high cultural and linguistic demand. However, significant differences in scores between referred samples and non-referred samples, regardless of linguistic or cultural characteristics, is not uncommon (Canivez & Watkins, 1998; Watkins, 2010) and significant mean score differences between samples of Ells and monolingual English speakers on standardized IQ tests have not been identified consistently across studies investigating the validity of the C-LIM (Kranzler et al., 2010; Tychanska, 2009; Verderosa, 2007). Kranzler et al. (2010) reported that mean scores from their sample of students receiving English as a second language services who were administered a WJ-III Cog *did* follow Flanagan et al.'s hypothesized different profile. However, Verderosa (2007) found that mean scores from her sample of bilingual preschoolers who were administered the DAS *did not* follow Flanagan et al.'s predicted different score pattern profile and results from Tychanska (2009) indicated that scores from English speakers with Speech and Language disabilities and scores from Ells with SLD *did* follow Flanagan et al.'s different score pattern profile on the Wechsler scales, which do not support Flanagan et al.'s hypotheses.

Most importantly, the existence of mean score differences does not provide sufficient evidence to support the existence of individual differences (Watkins, 2009). Flanagan et al. (2007) claimed that, "cultural and linguistic differences serve to artificially depress the scores of diverse individuals" (p. 175-176) and that when subtest scores follow their hypothesized different profile,

"practitioners should interpret such results as being a reflection of cultural or linguistic differences rather than of true ability" (p. 181). Results of the present study indicated that the percent of Ells whose scores followed Flanagan et al.'s hypothesized different profile ranged from 0% to 8.2% for Ells designated not SLD for each SLD criteria used in the sensitivity analysis, from 7.7% to 10.4% for Ells designated SLD for each SLD criteria used in the sensitivity analysis, and 10.5% for Ells altogether. Kranzler et al. (2010) was the only study that also reported individual frequencies for the different profile and they reported that only 37% of individual participant scores from their sample of students receiving English as a second language services who were administered the WJ-III Cog followed the hypothesized different profile pattern.

A more appropriate method for determining the clinical utility of the C-LIM is ROC analysis (Pepe, 2003; Swets, Dawes, & Monahan, 2000). Results of the ROC analyses from the present study indicated that the C-LIM produced similar decisions for scores from Ells with SLD and scores from Ells without SLD when they were compared to scores from the WISC-IV normative sample. This indicates that Flanagan et al.'s (2007) claim that the absence of the different profile in scores from students who are culturally and linguistically diverse is likely due to, "learning disability because the results will vary more a function of which test scores happen to be low and what construct those tests are designed to measure" (p. 197) is unsubstantiated. Results of the ROC analyses conducted between: (a) Ells with SLD and the WISC-IV normative sample, (b) Ells without SLD and the WISC-IV normative sample, and (c) Ells with SLD and Ells without SLD indicated that the C-LIM would be able to produce accurate decisions for a randomly chosen pair of participants from each subgroup around 50% of the time. Additionally, these results remained consistent when the criteria used to define SLD were manipulated. This means that the accuracy of the C-LIM decisions is not affected by how SLD is defined. Table 6 summarizes the study hypotheses and results.

Table 6

Hypotheses and Results for Each Comparison Group Disaggregated by SLD Criteria

	Different[a]	Disordered[b]	Hypotheses	Results
SLD criteria	*n*	*n*	AUC	AUC [95% CI]
Ells with SLD vs. WISC-IV normative sample				
MET decision	76	2,033	0.50-0.70	.52 [0.45, 0.59]
Regression discrepancy	55	2,033	0.50-0.70	.52 [0.44, 0.60]
One standard deviation discrepancy	48	2,033	0.50-0.70	.53 [0.44, 0.61]
Two standard deviation discrepancy	13	2,033	0.50-0.70	.51 [0.35, 0.67]
All Ells	86	2,033	0.50-0.70	.53 [0.46, 0.59]
Ells without SLD vs. WISC-IV normative sample				
MET decision	10	2,033	0.50-0.70	.53 [0.35, 0.70]
Regression discrepancy	31	2,033	0.50-0.70	.49 [0.39, 0.60]
One standard deviation discrepancy	38	2,033	0.50-0.70	.50 [0.41, 0.59]
Two standard deviation discrepancy	73	2,033	0.50-0.70	.48 [0.41, 0.55]
All Ells	86	2,033	0.50-0.70	.53 [0.46, 0.59]
Ells with SLD vs. Ells without SLD				
MET decision	10	76	0.50-0.70	.55 [0.37, 0.72]
Regression discrepancy	31	55	0.50-0.70	.49 [0.36, 0.62]
One standard deviation discrepancy	38	48	0.50-0.70	.53 [0.40, 0.65]
Two standard deviation discrepancy	73	13	0.50-0.70	.50 [0.33, 0.66]

Note. MET = Multidisciplinary Evaluation Team; SLD = Specific Learning Disability; AUC = Area Under the Curve; CI = Confidence Interval; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition.

[a]This refers to the true state of participants in each subgroup who are considered to be culturally and linguistically different from the WISC-IV normative sample.

[b]This refers to the true state of participants in each subgroup who are considered to be culturally and linguistically similar to the WISC-IV normative sample

Limitations

As with all applied research, the present study is not without limitations. The referred sample size was small and small sample sizes inflate type I and type II error. Type I error occurs when results of a statistical test indicate that there is a statistical difference in a study sample, but no difference exists in the population as a whole and type II error occurs when results of a statistical test indicate that there is not a statistical difference in a study sample, but there is a difference in the population. The sample sizes for Ells with SLD ranged from 13 to 76 across all 4 SLD criteria and the sample sizes for Ells without SLD in the present study ranged from 10 to 73 across all 4 SLD

criteria. Results of a power analysis indicated that 87 participants would be needed in each subgroup in order to detect an AUC of 0.59 or higher. AUC values for the present study all fell below 0.59. Consequently, the small size of the referred sample did not likely influence results.

A second limitation includes the absence of matching criteria on characteristics that may have influenced score variation such as age, gender, FSIQ score, and type of achievement test used to determine the presence of SLD. WISC-IV IQ scores are based upon age normative groups (Wechsler, 2003). This means that IQ scores are obtained by comparing the performance of an examinee to the performance of other individuals who are the same age as the examinee. As a result, age differences amongst participant subgroups may have affected scores. Gender is another variable that affects IQ scores. Empirical evidence suggests that differences exist between females and males on verbal abilities, quantitative abilities, and visual-spatial abilities (Macoby & Jacklin, 1974), although current estimates suggest that these differences are small (Janet, 1981). Variation between participant subgroups on FSIQ scores may have also affected score variation because there is some evidence to suggest that the general cognitive ability factor (e.g., Full Scale IQ score for the WISC-IV) is stronger for individuals who have lower cognitive ability compared to individuals who have higher cognitive ability (Abad, Colom, Juan-Espinosa, & Garcia, 2003; Deary, Egan, Gibson, Austin, Brand, & Kellaghan, 1996; Detterman & Daniel, 1989; Der & Deary, 2003). This phenomenon, commonly referred to as Spearman's law of diminishing returns, results in greater subtest scatter amongst the scores of individuals with higher general cognitive ability compared to the subtest scatter amongst the scores of individuals with lower general cognitive ability (Evans, 1999; Jensen, 2003; Legree, Pifer, & Grafton, 1996; te Nijenjuis & Hartmann, 2006). Lastly, referred participants were given the WJ-III or the WIAT-III as part of their special education evaluation. There may be specific reasons why a practitioner chooses to administer the WJ-III Ach instead of the WIAT-III and these reasons may be tied to the hypothesized cultural and linguistic demands of the subtests contained within each respective achievement test. Correlations between scores on the WJ-III Ach and the WIAT-III thought to measure similar constructs range from .31 to .82 (McGrew, Schrank, & Woodcock, 2007).

Finally, the reliance on archival data limited the amount of information that was gathered for the analyses. Practitioners are permitted to choose the test battery that is administered in special

education evaluations and there may be reasons why certain tests are chosen over others. Furthermore, these reasons may be tied to considerations regarding the cultural and/or linguistic background of the examinee. The WIAT-III and WJ-III were used to determine the presence of SLD in the sensitivity analysis and these issues may have influenced the frequency with which either test was administered.

Future research should continue to address the clinical utility of the C-LIM. The present study only investigated the ability of the WISC-IV C-LIM to produce accurate decisions, but there are 19 other standardized IQ tests for which Flanagan et al. (2007) have created C-LIMs that have yet to be investigated in this manner. It is essential that future research continue to use appropriate statistical methods for determining the clinical utility of the C-LIM as a diagnostic tool, such as ROC analyses. Mean score differences do not necessarily substantiate diagnosis of individual differences (Watkins, 2009). This research is necessary because multidisciplinary evaluation teams use SLD diagnostic decisions to allocate special education resources to students and the appropriate allocation of resources will likely increase the frequency of positive student outcomes.

Conclusions

The results of the present study should be considered preliminary due to its several limitations. However, it is not recommended that the C-LIM be used in applied practice at this time. Flanagan et al. (2007) claimed the C-LIM, "may well prove to be of significant practical value in decreasing bias related to the selection and interpretation of tests" (p. 175). This claim is currently unsubstantiated, and the results of the present investigation dispute its verity. Results indicated that the different profile did not exist amongst the scores from Ells with and without SLD regardless of the way in which SLD was defined. Further, the presence of this pattern did not accurately discriminate between participants who were Ells with SLD and Ells who were without SLD or participants from the WISC-IV normative sample. The lack of discriminatory power associated with the different profile pattern implies that the WISC-IV subtests are not accurately classified as having low, moderate, or high cultural and/or linguistic demands and it calls into question the relevancy of gathering such information in the diagnosis of SLD.

Similar conclusions have been made about other attempts to ascribe meaning to IQ subtest score patterns, commonly referred to as "profile analysis." Over a decade ago, Bray, Kehle, and Hintze (1998) suggested that the reason profile analysis persists may be because, "the notion that a single IQ score captures all that is meaningful and practical about the IQ test is simply not acceptable, regardless of evidence to the contrary" (p. 209). Others have posited that profile analysis continues to be advanced in spite of unsubstantiated evidence because of intuitive appeal and an over reliance on clinical judgment (Garb, 2003). Watkins, Glutting, and Youngstrom (2005) cautioned that, "scientific psychological practice cannot be substantiated by clinical conjectures, personal anecdotes, and unverifiable beliefs that have consistently failed empirical validation" (p. 263).

Moreover, the APA Ethics Code (2002) states that, "psychologists use assessment instruments whose validity and reliability have been *established* [emphasis added] for use with members of the population tested" and "when such validity or reliability has not been established, psychologists describe the strengths and limitations of test results and interpretation" (Section 9.02(b). The Standards for Educational and Psychological Testing (1999) also provide guidelines for the use and validation of new test instruments. Standards 1.4 and 1.12 state that, "if a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary" and that, "when interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided." However, the results of Kranzler et al. (2010) and those of the present study indicate that the validity and reliability of the C-LIM are poor, which places the burden of proof to substantiate the use of the C-LIM on the practitioners who choose to use it. The C-LIM has intuitive appeal; however, current evidence does not support its use as a diagnostic tool.

References

Abad, F. J., Colom, R., Juan-Espinosa, M., & Garcia, L. F. (2003). Intelligence differentiation in adult samples. *Intelligence, 31,* 157-166.

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

American Psychological Association (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060-1073. doi: 10.1037//0003-066X.57.12.1060

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

American Psychological Association (2000). *Report of the task force on test user qualifications.* Retrieved from http://www.apa.org/science/programs/testing/test-clearinghouse.aspx

Anonymous. (N.D.) *LD technical supplement.*

Aud, S., Hussar, W., Planty, M., Snyder, T., Bianco, K., Fox, M., … Hannes, G. (2010). *The Condition of Education 2010* (NCES 2010-028). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Bailey, N. (1949). Consistency and variability in the growth of intelligence from birth to eighteen years. *Journal of Genetic Psychology, 75,* 165-196.

Baldessarini, R. J., Finkelstein, S., & Arana, G. W. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry, 40,* 569-573.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the Receiver Operating Characteristic graph. *Journal of Mathematical Psychology, 12,* 387-415.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48,* 5-37. doi: 10.1016/j.jsp.2009.10.001

Begeny, J. C., Laugle, K. M., Krouse, H. E., Lynn, A. E., Tayrose, M. P., & Stage, S. A. (2010). A control-group comparison of two reading fluency programs: The Helping Early Literacy With Practice Strategies (HELPS) program and the Great Leaps K-2 reading program. *School Psychology Review, 39,* 137-155.

Bradway, K. P., Thompson, C. W., & Cravens, R. B. (1958). Preschool IQ's after twenty-five years. *Journal of Educational Psychology, 49,* 278-281

Bradway, K. P., & Thompson, C. W. (1962). Intelligence at adulthood: A twenty-five year follow-up. *Journal of Educational Psychology, 53,* 1-14. doi: 10.1037/h0045764

Bray, M. A., Kehle, T. J., & Hintze, J. M. (1998). Profile analysis with the Wechsler scales: Why does it persist? *School Psychology International, 19,* 209-220. doi: 10.1177/0143034398193002

Breaux, K. C. (2009). *Wechsler Individual Achievement Tests-Third Edition technical manual with adult norms.* Bloomington, MN: Pearson.

Bruck, M. (1987). The adult outcomes of children with learning disabilities. *Annals of Dyslexia, 37,* 252-263.

Burns, M. K., Codding, R. S., Boice, C. H., & Lukito, G. (2010). Meta-analysis of acquisition and fluency math interventions with instructional and frustrational level skills: Evidence for a skill-by-treatment interaction. *School Psychology Review, 39,* 69-83.

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition. *Psychological Assessment, 10,* 285-291. doi: 1040-3590/98

Cantor, S. B., & Kattan, M. W. (2000). Determining the area under the ROC curve for a binary diagnostic tests. *Medical Decision Making, 20,* 468-470. doi: 10.1177/0272989X0002000410

Centor, R. M., & Schwartz, J. S. (1985). An evaluation of methods for estimating the Area Under the Receiver Operating Characteristic (ROC) curve. *Medical Decision Making, 5,* 149-158. doi: 10.1177/0272989X8500500204

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159. doi: 0033-2909/92

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cohn, N. (1961). Understanding the process of adjustment to disability. *Journal of Rehabilitation, 27,* 16-18.

Colarusso, R. P., Keel, M. C., & Dangel, H. L. (2001). A comparison of eligibility criteria and their impact on minority representation in LD programs. *Learning Disabilities Research and Practice, 16*, 1-7.

Cortiella, C. (2011). The state of learning disabilities. *National Center for Learning Disabilities.* Retrieved from http://www.ncld.org/stateofld

Data Accountability Center. (2009). *IDEA part B child count* [Data file]. Retrieved from www.IDEAdata.org

Data Accountability Center. (2008-2009). *IDEA part B exiting (2008-09)* [Data file]. Retrieved from www.IDEAdata.org

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95-106.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668-1674.

Deary, I. J., Egan, V., Gibson, G. J., Austin, E. J., Brand, C. R., & Kellaghan, T. (1996). Intelligence and the differentiation hypothesis. *Intelligence, 23,* 105-132.

Delaquadri, J., Greenwood, C. R., Stretton, K., & Hall, R. V. (1983). The peer tutoring game: A classroom procedure for increasing opportunity to respond and spelling performance. *Education and Treatment of Children, 6,* 225-239.

Der, G., & Deary, I. J. (2003). IQ, reaction time, and the differentiation hypothesis. *Intelligence, 31,* 491-503.

Detterman, D. K., & Daniel, M. H. (1989). Correlates of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence, 13,* 349-359.

Dhaniram-Beharry, E. (2008). Cultural and linguistic influences on test performance: Evaluation of alternative variables (Doctoral dissertation). Retrieved from ProQuest (Accession No. 3336081)

Dorfman, D. D., & Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating-method data. *Journal of Mathematical Psychology, 6,* 487-496. doi: 480/6/3-11

Dynda, A. M. (2008). The relation between English language proficiency and IQ test performance (Doctoral dissertation. Retrieved from ProQuest. (Accession No. 3340910)

Elwood, R. W. (1993). Psychological tests and clinical discriminations: Beginning to address the base rate problem. *Clinical Psychology Review, 13,* 409-419. doi: 0272-7358/93

Evans, M. G. (1999). On the asymmetry of g. *Psychological Reports, 85,* 1059-1069.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 2,* 861-874.

Figueroa, R. A., Newsome, P. (2004). The diagnosis of LD in English learners: Is it nondiscriminatory? *Journal of Learning Disabilities, 39,* 206-214.

Finucci, J. M. (1986). Follow-up studies of developmental dyslexia and other learning disabilities. In S. D. Smith (Ed.), *Genetics and learning disabilities* (pp. 97-121). San Diego, CA: College Hill Press.

Finucci, J. M., Gottfredson, L. S., & Childs, B. (1984). Explaining the adult careers of dyslexic boys: Variations in critical skills for high level jobs. *Journal of Vocational Behaviors, 24,* 355-373.

Flanagan, D. P., & Otiz, S. O. (2001). How to apply CHC cross-battery assessment to culturally and linguistically diverse individuals. In A. S. Kaufman & N. L. Kaufman (Series Ed.), *Essentials of cross-battery assessment* (pp. 213-270). Hoboken, NJ: John Wiley and Sons.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). Use of the cross-battery approach in the assessment of diverse individuals. In A. S. Kaufman & N. L. Kaufman (Series Ed.), *Essentials of cross-battery assessment second edition* (2nd ed., pp. 146-205). Hoboken, NJ: John Wiley and Sons.

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of IQs: Implications for professional psychology. *Professional Psychology: Research and Practice, 39,* 414-423.

Fulk, B. M., & Stormont-Spurgin, M. (1995). Spelling interventions for students with disabilities: A review. *The Journal of Special Education, 28,* 488-513.

Garb, H. N. (2003). Clinical judgment and mechanical prediction. In I. B. Weiner (Series Ed.) & J. R. Graham & J. A. Naglieri (Vol. Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 27-42). New York: Wiley.

Garfield, S. L. (1978). Research problems in clinical diagnosis. *Journal of Consulting and Clinical Psychology, 46,* 596-607. doi: 10.1037/0022-006X.46.4.596

Gerardi, R., Keane, T. M., & Penk, W. (1989). Utility: Sensitivity and specificity in developing diagnostic tests of combat-related post-traumatic stress disorder (PTSD). *Journal of Clinical Psychology, 45,* 691-703.

Gettinger, M., & Kosick, R. (2001). Psychological services for children with disabilities. In J. N. Hughes, A. M. LaGreca, & J. C. Conoley (Eds.), *Handbook of psychological services for children and adolescents* (pp. 421-435). New York, NY: Oxford University Press.

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

Gottesman, R. L. (1979). Follow-up of learning disabled children. *Learning Disability Quarterly, 2,* 60-68.

Gottesman, R. L. (1989). The adult with learning disabilities: An overview. *Learning Disabilities, 5,* 1-14.

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools – A report to Carnegie Corporation of New York.* Washington, DC: Alliance for Excellent Education.

Griner, P. F., Mayewski, R. J., Mushlin, A. L., & Greenland, P. (1981). Selection and interpretation of diagnostic tests and procedures. *Archives of Internal Medicine, 94,* 557-592.

Hambleton, R. K. & Li, S. (2005). Translation and adaptation issues and methods for educational and psychological tests. In C. L. Frisby & C. R. Reynolds (Eds.), *Comprehensive handbook of multicultural school psychology* (pp. 881-903). Hoboken, NJ: John Wiley and Sons.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology, 143,* 29-36.

Harring, K. A., Lovett, D. L., & Smith, D. D. (1990). A follow-up study of recent special education graduates of learning disabilities programs. *Journal of Learning Disabilities, 23,* 108-113.

Higgins, E. L., Raskind, M. H., Goldberg, R. J., & Herman, K. L. (2002). Stages of acceptance of a learning disability: The impact of labeling. *Learning Disabilities Quarterly, 25,* 3-18.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1400

Janet, H. S. (1981). How large are cognitive gender differences? A meta-analysis using w^2 and d. *American Psychologist, 36,* 892-901. doi: 10.1037/0003-066x.36.8.892

Jensen, A. R. (2003). Regularities in Spearman's law of diminishing returns. *Intelligence, 31,* 95-105.

Joint Committee of Testing Practices (2004). *Code of fair testing practices in education.* Retrieved from http://www.apa.org/science/programs/testing/fair-code.aspx

Kaufman, A. S., & Lichtenberber, E. O. (2000). *Essentials of WISC-III and WPPSI-R assessment.* New York, NY: Wiley.

Kessel, J. B., & Zimmerman, M. (1993). Reporting errors in studies of the diagnostic performance of self-administered questionnaires: Extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. *Psychological Assessment, 5,* 395-399.

Kopp, K. H., Miller, J. H., & Mulkey, S. W. (1984). The paradox of learning disabilities: A stumbling bloc to rehabilitation. *Journal of Rehabilitation, 50,* 4-5.

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

Kurzweil, S. R. (1992). Developmental reading disorder: Predictors of outcome in adolescents who received early diagnosis and treatment. *Developmental and Behavioral Pediatrics, 13,* 399-404.

Legree, P. J., Pifer, M. E., Grafton, F. C. (1996). Correlations among cognitive abilities are lower for higher groups. *Intelligence, 23,* 45-57.

Macoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Paalo Alto CA: Stanford University Press.

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

Massetti, G. M. (2009). Enhancing emergent literacy skills of preschoolers from low-income environments through a classroom-based approach. *School Psychology Review, 38,* 554-569.

McDermott, P. A., Goldberg, M. M., Watkins, M. W., Stanley, J. L., & Glutting, J. J. (2006). A nationwide epidemiologic modeling study of LD: Risk, protection, and unintended impact. *Journal of Learning Disabilities, 39,* 230-251.

McDermott, R., & Varenne, H. (1995). Culture as disability. *Anthropology and Education Quarterly, 26,* 324-348.

McGrew, K. S., & Woodcock, R. W. (2001). Technical Manual. *Woodcock-Johnson III.* Itasca, IL: Riverside Publishing.

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

Meehl, P. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, MN: University of Minnesota.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194-216.

Menesses, K. F., & Gresham, F. M. (2009). Relative efficacy of reciprocal and nonreciprocal peer tutoring for student at-risk for academic failure. *School Psychology Quarterly, 24,* 266-275. doi: 10.1037/a0018174

Mercer, C. D., Jordan, L., Allsopp, D. H., & Mercer, A. R. (1996). Learning disabilities definitions and criteria used by state education departments. *Learning Disabilities Quarterly, 19,* 217-232.

Messick, S. (1995). Validity of psychological assessment validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8,* 283-298.

Moldin, S. O., Gottesman, I. B., Rice, J. P., & Erlenmeyer-Kimling, L. (1991). Replicated psychometric correlates of schizophrenia. *American Journal of Psychiatry, 148,* 762-767.

National Association of School Psychologists (2010). National Association of School Psychologists principles for professional ethics. *School Psychology Review, 39,* 302-319.

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 (Doctoral dissertation). Retrieved from ProQuest. (Accession No. 3286026)

te Nijenjuis, J., & Hartmann, P. (2006). Spearman's "law of diminishing returns" in samples of Netherlands and immigrant children and adults. *Intelligence, 34,* 437-447.

Novick, M. R. (1966). The axioms and principal results of classical test theory *Journal of Mathematical Psychology, 3*, 1-18. doi: 10.1016/0022-2496(66)90002-2

Ochoa, S. H., Riccio, C., Jimenez, S., Garcia de Alba, R., & Sines, M. (2004). Psychological assessment of ELLs and/or bilingual students: An investigation of school psychologists current practices. *Journal of Psychoeducational Assessment, 22*, 185-208. doi: 10.1177/073428290402200301

Olin, J. T., Schneider, L. S., Eaton, E. M., Zemansky, M. F., & Pollock, V. E. (1992). The Geriatric Depression Scale and the Beck Depression Inventory as screening instruments in an older adult outpatient population. *Psychological Assessment, 4,* 190-192. doi: 1040-3590/92

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

Ortiz, S. O., & Ochoa, S. H. (2005). Advances in cognitive assessment of culturally and linguistically diverse individuals: A nondiscriminatory interpretive approach. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 234-250). New York, NY: Guilford.

Pepe, M. S. (2003). *Statistical evaluation of medical tests for classification and prediction.* New York, NY: Oxford University Press.

Pintea, S., & Moldovan, R. (2009). The receiver-operating characteristics (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Cognitive and Behavioral Psychotherapies, 9*, 49-66.

Rapp, S. R., Parisi, S. A., Walsh, D. A., & Wallace, C. E. (1988). Detecting depression in elderly medical inpatients. *Journal of Consulting and Clinical Psychology, 56,* 509-513.

Raskind, M. H., Goldberg, R. J., Higgins, E. L., & Herman, K. L. (1999). Patterns of change and predictors of success in individuals with learning disabilities: Results from a twenty-year longitudinal study. *Learning Disabilities Research and Practice, 14,* 35-49.

Reiff, H. B., Gerber, P. J., & Ginsberg, R. (1997). *Succeeding expectations: Successful adults with learning disabilities.* Austin, TX: Pro-ed.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2006). *Measurement and assessment in education.* Boston, MA: Pearson Education, Inc.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior, 29*, 615-620. doi: 10.1007/s10979-005-6832-7

Rogan, L. L. & Hartman, L. D. (1990). Adult outcome of learning disabled students ten years after initial follow-up. *Learning Disabilities Focus, 5,* 91-102.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in the psychological science. *American Psychologist, 44,* 1276-1284.

Sanchez, S. (1984). Where do we go from here: A look to the future in the rehabilitation of learning disabled persons. *Journal of Rehabilitation, 50,* 82-88.

Schonhaut, S., & Satz, P. (1983). Prognosis for children with learning disabilities: A review of follow-up studies. In M. Rutter (Ed.), *Developmental neuropsychiatry* (pp. 542-563). New York, NY: Guilford Press.

Shapiro, J. H. (1999). Bounds on the area under the ROC curve. *Journal of the Optical Society of America, 16,* 53-57. doi: 10.1364/JOSAA.16.000053

Shepard, L. (1980). An evaluation of the regression discrepancy method for identifying children with learning disabilities. *Journal of Special Education, 14,* 79-91. doi: 10.1177/002246698001400108

Sitlington, P. L., & Frank, A. R. (1990). Are adolescents with learning disabilities successfully crossing the bridge into adult life? *Learning Disability Quarterly, 13,* 97-111.

Sotelo-Dynega, M. (2007). *Cognitive performance and the development of English language proficiency* (Doctoral dissertation). Retrieved from ProQuest. (Accession No. 3282715)

Spekman, N.J., Goldberg, R. J., & Herman, K. L (1992). Learning disabled children grow up: A search for factors related to success in the young adult years. *Learning Disabilities Research and Practice, 7,* 161-170.

Spreen, (1988). *Learning disabled children growing up: A follow-up into adulthood.* New York,NY: Oxford University Press.

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to Receiver Operating Characteristics curves. *The Canadian Journal of Psychiatry, 52,* 121-128.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1-26.

Tychanska, J. (2009). Evaluation of speech and language impairment using the culture-language test classification and interpretive matrix (Doctoral dissertation). Retrieved from ProQuest. (Accession No. 3365687)

U.S. Department of Education, Office of Special Education and Rehabilitative Services, Office of Special Education Programs. (2005). *26th Annual (2004) Report to Congress on the Implementation of the Individuals with Disabilities Education Act, vol. 1,* Washington, D.C.

Verderosa, F. A. (2007). Examining the effects of language and culture on the differential ability scales with bilinguial preschoolers (Doctoral dissertation). Retreived from ProQuest. (Accession No. 3286027)

Wagner, M., Marder, C., Blackorby, J., Cameto, R., Newman, L., Levine, P., & Davies-Mercier, E. (2003). *The achievements of youth with disabilities during secondary school. A report from the National Longitudinal Transition Study-2.* Menlo Park, CA: SRI International.

Wagner, M., Newman, L., Cameto, R., Garza, N., & Levine, P. (2005). *After high school: A first look at the postschool experiences of youth with disabilities. A report from the National Longitudinal Transition Study-2 (NLTS2).* Menlo Park, CA: SRI International.

Watkins, M. W. (2003). IQ subtest analysis: Clinical acumen or clinical illusion. *Scientific Review of Mental Health Practice, 2,* 118-141.

Watkins, M. W. (2010). Structure of the Wechsler Intelligence Scale for Children-Fourth Edition Among a National Sample of Referred Students. *Psychological Assessment, 22,* 782-787. doi: 10.1037/a0020043

Watkins, M. W., Glutting, J. J., & Youngstrom, E. A. (2005). Issues in subtest profile analysis. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 251-268). New York, NY: Guilford.

Wechsler, D. (2003a). *Wechsler Intelligence Scale for Children-Fourth Edition.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003b). *WISC-IV technical and interpretive manual.* San Antonio, TX: The Psychological Corporation.

Weiss, L. G., Beal, A. L., Saklofske, D. H., Alloway, T. P., & Prifitera, A. (2008). Interpretation and intervention with the WISC-IV in the clinical assessment context. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical assessment and intervention* (pp. 3-66). San Diego, CA: Academic Press.

Werner, E. E. (1993). Risk and resilience in individuals with learning disabilities: Lessons learned from the Kauai longitudinal study. *Learning Disabilities Research and Practice, 8,* 28-35.

Zehler, A. M., Fletschman, H. F., Hopstock, P. J., Stephenson, T. G., Pendzic, M. L., & Sapru, S. (2003). *Descriptive study of services to LEP students and LEP students with disabilities. Volume 1: Research Report.* Arlington, VA: Development Associates. Retrieved from http://www.ncela.gwu.e