School Psychology Review,
2013, Volume 42, No. 4, pp. 367–382

Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students

Kara M. Styck
Arizona State University

Marley W. Watkins
Baylor University

Abstract. The Culture-Language Interpretive Matrix (C-LIM) was developed by Flanagan, Ortiz, and Alfonso (2013) to evaluate the extent to which developmental language proficiency and acculturative learning opportunity affect the validity of standardized test scores for individual students. According to this approach, validity may be compromised for children with cultural and language experiences, such as learning English as a second language, that differ from the population on which the test was normed. This study employed diagnostic utility statistics to test whether the C-LIM for the Wechsler Intelligence Scales for Children—Fourth Edition (WISC-IV) could accurately distinguish between students from a referred sample of English language learners (n = 86) and monolingual students without disabilities from the WISC-IV normative sample (n = 2,033). Results indicated that the C-LIM identified the English language learner children at chance levels. Evidence from previous studies as well as the current negative results does not support use of the C-LIM for making decisions about individual students.

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). Construct validity is especially important in high-stakes decisions, such as in the determination of eligibility for special education services. Special education eligibility decisions often rely on interpretations of standardized norm-referenced IQ test scores such as in the diagnosis of specific learning disabilities (SLD; Mercer, Jordan, Allsopp, & Mercer, 1996). Norm-referenced scores are interpreted by comparing the performance of an examinee to a normative comparison group (Reynolds, Livingston, & Willson, 2006). However, the construct va-

This study is based on the dissertation of the first author. The contributions of Dr. Roy Levy and Dr. John Balles are gratefully acknowledged. Standardization data from the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV). Copyright © 2003 NCS Pearson, Inc. Used with permission. All rights reserved.

Kara Styck is now at the University of Texas—San Antonio.

Correspondence concerning this article should be addressed to Marley W. Watkins, Baylor University, Department of Educational Psychology, One Bear Place #97301, Waco, TX 76798-7301. email: Marley_Watkins@baylor.edu

lidity of test scores may be compromised when the examinee and the normative comparison group do not share similar demographic and experiential characteristics (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

The cultural and linguistic makeup of the U.S. school-aged population is becoming increasingly diverse (Aud et al., 2010) and variance from test items that reflects other variables such as cultural and linguistic background may invalidate score interpretations. The Individuals with Disabilities in Education Act (2004) requires that states have in place "policies and procedures designed to prevent the inappropriate over identification 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 are disproportionately at risk for being identified as having SLD (U.S. Department of Education, 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).

Some researchers believe that cultural and linguistic diversity depresses IQ scores, which contributes to the disproportionate identification of these students for special education programming (Flanagan, Ortiz, & Alfonso, 2007, 2013; Ortiz, 2011). For instance, Flanagan et al. (2013) recently asserted that test score interpretations are invalid for "individuals whose experiences and development (not from intrinsic delays but rather circumstantial changes as might occur with respect to learning English as a second language) differ from those… established by the population on whom the test was normed" (p. 294). Following this logic, invalid test scores would lead to erroneous special education placement for students who are learning English as a second language.

The Culture-Language Interpretive Matrix

To address these issues, Flanagan et al. (2007) created the Culture-Language Interpretive Matrix (C- LIM). The C-LIM is a 3 by 3 matrix that is hypothesized to help practitioners understand the test scores of individual students and directly assess whether assessment data were primarily influenced by cultural or linguistic variables (Flanagan et al., 2013). If true, the test scores would be invalid and should not be interpreted because they reflect cultural and linguistic experiential differences between the examinee and the children in the normative sample. If false, the test scores are considered valid and should be interpreted. This phenomenon was referred to as "difference versus disorder" (Flanagan et al., 2007, p. 175), and provides a yes/no decision regarding the validity of interpretations made from test scores as a result of the hypothesized construct-irrelevant variance from cultural and linguistic influences. Thus, the purported purpose of the C-LIM is to assist in evaluating the validity of standardized testing results and to determine if they are interpretable or not (Flanagan et al., 2013).

Figure 1 portrays a generic C-LIM. 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 standardized intelligence tests that were classified by Flanagan et al. (2007, 2013) as having low, moderate, or high cultural and linguistic demand. Each C-LIM includes the subtests of a single test battery. For example, there is a C-LIM to accompany the Stanford-Binet Intelligence Scales—Fifth Edition (Roid, 2003) and a separate C-LIM to accompany the Woodcock-Johnson Tests of Cognitive Abilities—Third Edition (WJ III Cog; Woodcock, McGrew, & Mather, 2001).

The C-LIM analysis is conducted through three sequential steps. First, subtest scaled scores are converted to a standard metric (M = 100; SD = 15) and inserted into the appropriate cells of the matrix. Second, the mean of each cell is computed. Lastly, the cell

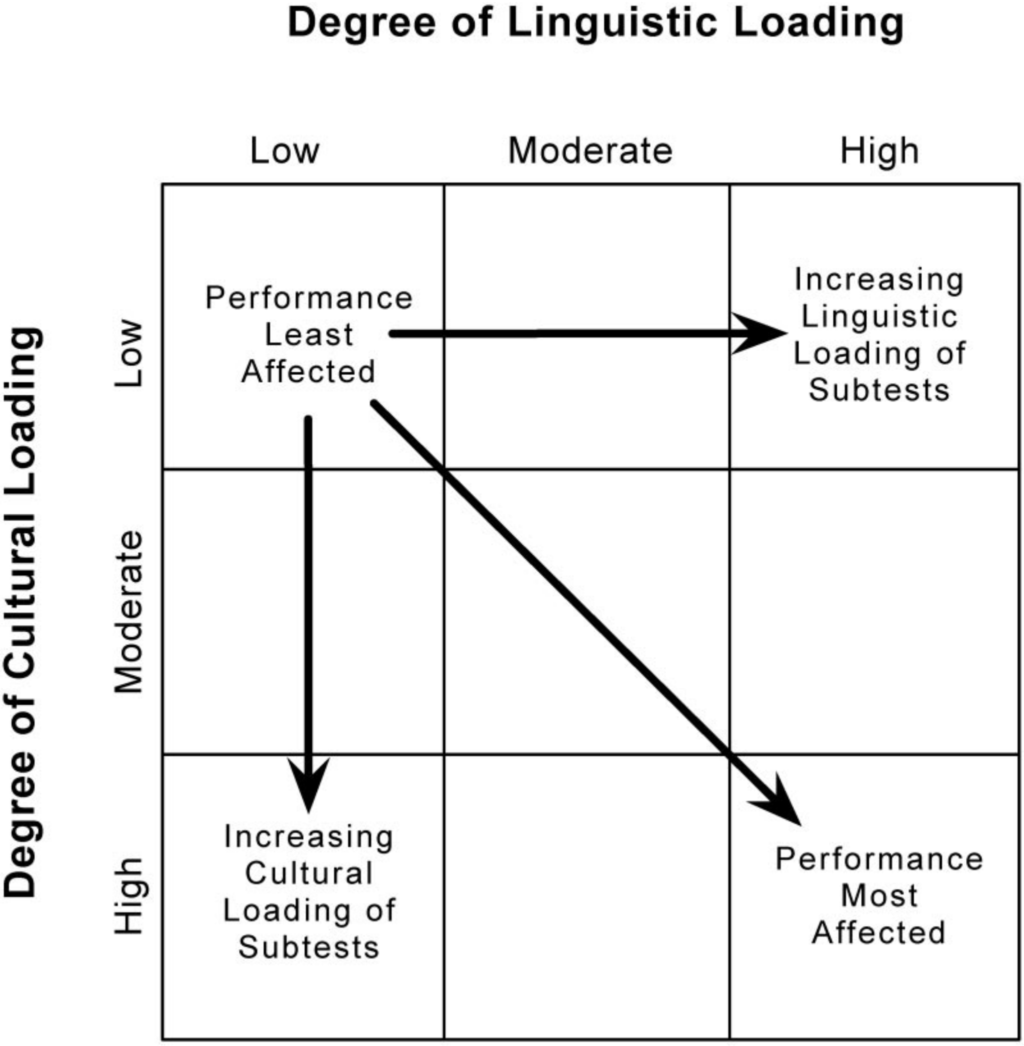

Degree of Linguistic Loading
Low
Moderate
High
Degree of Cultural Loading
Low
Moderate
High
Performance Least Affected
Increasing Linguistic Loading of Subtests
Increasing Cultural Loading of Subtests
Performance Most Affected

Figure 1. A generic Culture-Language Interpretive Matrix.

means along the diagonal of the matrix are examined for the presence of a profile that meets the following criteria: (a) the upper left-hand corner of the matrix contains the highest cell mean, (b) the lower right-hand corner of the matrix contains the lowest cell mean, and (c) the means decline down the diagonal of the matrix from the upper left-hand corner down to the lower right-hand corner. For test scores that follow this pattern, Flanagan and colleagues (2007) recommended that practitioners recognize the invalidity of assessment data and to not interpret them for individual students. In contrast, scores from students designated as valid would not match any specific pattern and would not result from experiential differences (Ortiz, 2011). Taken together, this suggests that the declining profile should only emerge when individuals are culturally and linguistically different from the normative group of a test (i.e., scores considered invalid by the C-LIM) and that the absence of the declining profile should only emerge when individuals share cultural and linguistic experiences with the normative group of a test (i.e., scores considered valid by the C-LIM). By definition, children in the normative group are the standard and should exhibit a valid C-LIM profile.

The C-LIM has been presented as a valuable tool to aid in the comprehensive evaluation of culturally and linguistically diverse students (Flanagan et al., 2007, 2013; Ortiz, 2011). However, there are several potential concerns with the C-LIM. First, the C-LIM categorized subtests from tests of intelligence as having low, medium, or high cultural and linguistic demand without regard for empirical evidence regarding the reliability and validity of those subtests. Second, converting continuous constructs like culture and language into a few categories may result in a loss of information and biased measurement (Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002). Third, classification of most of the IQ test batteries was accomplished with an expert consensus procedure (Flanagan et al., 2007), which suggests that clinical judgment was the primary method used to classify numerous IQ subtests as having low, moderate, or high linguistic and cultural demand. Diagnostic decisions based on clinical judgment can be inherently challenging (Weiss, Shanteau, & Harries, 2006) and less accurate than diagnostic decisions based on actuarial methods (Dawes & Corrigan, 1974; Dawes, Faust, & Meehl, 1989; Meehl, 1954).

Historically, subjective judgments of the cultural loading of cognitive test items have been inaccurate (Jensen, 1976; Sandoval & Miille, 1980) and empirical evidence suggests that the C-LIM IQ subtest classifications of low, moderate, and high cultural and linguistic demand are similarly dubious (Cormier, 2012; Cormier, McGrew, & Evans, 2011). For example, Cormier (2012) advised that the C-LIM not be used in test interpretation because nearly half of the subtests classified within the WJ-III Cog C-LIM did not account for a proportion of subtest score variance that corresponded appropriately to Flanagan et al.'s (2007, 2013) low, moderate, and high linguistic and cultural demand classifications in a series of structural equation models. Moreover, the assumption that systematic bias results from minority representation in the normative sample was not empirically supported (Fan, Willson, & Kapes, 1996).

Finally, empirical evidence regarding the criterion validity of C-LIM profile interpretations is sparse and inconsistent. Only eight studies that directly assessed the criterion validity of the C-LIM have been conducted to date and only one of those eight (i.e., Kranzler, Flores, & Coady, 2010) has been published in a peer-reviewed journal. Kranzler et al. (2010) reported that mean scores from their sample of students receiving English as a second language services for limited English proficiency exhibited Flanagan et al.'s (2007, 2013) invalid profile of declining cell means for the WJ-III Cog C-LIM. However, they noted that only 37% of individual participant scores followed the invalid profile, whereas English was not the first language of 100% of their participants and 74% of their participants had resided in the United States for fewer than 3 years.

The remaining studies consist of unpublished doctoral dissertations, most of which were conducted under the direction of one of the C-LIM authors (Aziz, 2011; Brown, 2008; Dhaniram-Beharry, 2008; Nieves-Brull, 2006; Souravlis, 2010; Tychanska, 2009; Verderosa, 2007) and few of which reported results in favor of the C-LIM. For example, Verderosa (2007) reported that mean scores from a sample of bilingual preschoolers on the Differential Abilities Scales (DAS; Elliot, 1990) did not follow Flanagan et al.'s (2007) predicted invalid profile of decline, and Brown (2008) reported mean scores from a sample of ELLs on the Battería—Third Edition followed the declining profile for an unpublished C-LIM (Battería-III; Woodcock, Muñoz-Sandoval, McGrew, & Mather, 2007). Flanagan et al. (2007) specify that the clear declining pattern of some subgroups of ELLs may also indicate the presence of a disability, including ELL with speech and language impairment, developmental disabilities, or general cognitive impairment. However, mean scores from samples of ELLs identified with one of these disabilities in Aziz (2011), Souravlis (2010), and Tychanska (2009) did not follow the hypothesized pattern of decline. Aziz (2011) and Nieves-Brull (2006) remain the only dissertations that reported partial results in favor of the C-LIM. Mean scores from ELL with developmental disabilities in Aziz (2011) followed the hypothesized declining pattern and mean scores from ELL without disabilities in Nieves-Brull (2006) followed the hypothesized declining pattern. However, the base rate of ELLs Flanagan et al. (2007) expect to follow the pattern of decline remains unspecified.

Most importantly, no prior study has addressed the distinction between nomothetic and idiographic approaches to clinical decision making (Kraemer, Frank, & Kupfer, 2011; Weiner, 2003). Ortiz (2011) presented the C-LIM as an approach to assist practitioners in determining if test results are valid and interpretable or are primarily indicators of cultural and linguistic factors. Thus, the C-LIM was designed for making decisions about individuals (Flanagan et al., 2007, 2013), but most of the existing research on the C-LIM relied on statistical tests of mean differences between groups. Group differences are necessary but insufficient for making accurate decisions about individuals (Elwood, 1993) because group means reflect what happens to the typical member of the population sampled rather than any individual member of the population (Kraemer et al., 2011). Little attention has been paid to the score overlap between the groups, although its importance has long been known (Meehl & Rosen, 1955; Metz, 1978; Swets, Dawes, & Monahan, 2000). Decisions about individuals are more appropriately addressed with a diagnostic utility approach (Wiggins, 1987). Diagnostic utility statistics have been used extensively for the validation of educational, medical, and psychological tests (Jordan et al., 2010; Metz, 2008; Rapp, Parisi, Walsh, & Wallace, 1988), but have not been applied to the C-LIM.

Given that decisions based on the C-LIM may have a major impact on children, additional research on its accuracy is necessary. Consequently, the purpose of the present study was to address the following research questions:

1. What is the probability that the WISC-IV C-LIM can accurately distinguish between students who are culturally and linguistically diverse and students from the WISC-IV normative group?
2. What is the probability that a randomly selected culturally and linguistically diverse individual will have a higher C-LIM score than a randomly selected individual from the WISC-IV normative group?

It is beyond the scope of any single study to examine every C-LIM created by Flanagan et al. (2007). Therefore, the WISC-IV was selected because of its common use in applied settings (Kaufman & Lichtenberger, 2000) and its value as a diagnostic assessment instrument for children (Weiss, Beal, Saklofske, Alloway, & Prifitera, 2008).

Method

Participants

Two groups of participants were included in the present study: (a) ELLs and (b) monolingual English speakers. The ELL sample included 86 school-aged children (56 males, 30 females) aged 6–16 years (M = 11.3, SD = 2.4) who were referred for a special education evaluation in two Southwestern metropolitan school districts to determine initial or continued eligibility for special education services. Scores from approximately 91% (n = 78) of participants were obtained from initial evaluations and roughly 97% (n = 83) of participants were identified as meeting criteria for an educational disability (86% as SLD) by a multidisciplinary evaluation team. School records indicated that Spanish was the home language of all 86 participants and the primary language of 57% of the participants (n = 49).

The sample of monolingual English speakers was extracted from the WISC-IV normative sample. Permission was granted from National Computer Systems Pearson to use the WISC-IV normative sample in the present study. Participants included 2,033 students (1,004 males, 1,029 females) aged 6–16 years old (M = 11.5, SD = 3.2) who were administered the WISC-IV during the normative phase of test development and who were not diagnosed as having educational-related disabilities, although 19 (0.9%) of the participants were diagnosed as gifted and talented by independent practitioners.

Instruments

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 (FSIQ; *M* = 100; *SD* = 15; Wechsler, 2003a). The standardization sample included 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. Boys and girls were sampled in equal proportions, whereas the race/ethnicity of around 64% of the standardization sample was White, 16% African American, 15% Hispanic, 4% Asian, and 1% other. Students not fluent in English were specifically excluded from participation in the WISC-IV standardization process (Wechsler, 2003b). See Wechsler (2003b) for more detailed information.

Procedure

Following university institutional review board and school district approval, all of the active special education files for each school district were examined for the presence of WISC-IV scores and information was collected from files that contained these scores. Data that were collected included parent home language survey results, WISC-IV scores, and the disability eligibility status 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 invalid profile hypothesized to exist among scores from culturally and linguistically diverse students. 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, and (c) their file contained all 10 core WISC-IV subtest scores. Participants were excluded from both the referred sample and the WISC-IV normative sample if their WISC-IV FSIQ score was less than or equal to 73 to further ensure that students who met criteria for an intellectual disability were not included by accounting for the average standard error of measurement published in the WISC-IV technical manual (Wechsler, 2003b).

Analyses

Preliminary analyses were conducted to describe the WISC-IV scores obtained from study participants. Means and standard deviations for all WISC-IV subtest, composite, and FSIQ scores were computed. In addition, the degree to which WISC-IV subtest scaled scores, composite standard scores, and FSIQ scores differed at a statistically significant level between each of the participant subgroups was analyzed using a one-way analysis of variance (ANOVA). The Levene's test of homogeneity of variances was examined to test the degree to which the score variances between the participant subgroups were statistically significantly different and the Welch approximate *F* test was used to determine the overall statistical significance of the ANOVA because it does not assume homogeneity of variances. The alpha level for each individual test was set at .003 using the Bonferroni correction to maintain an experimentwise error rate of .05.

Diagnostic utility statistics were also computed. Following the criteria outlined by Flanagan et al. (2007, 2013), WISC-IV subtest scores for each participant were placed into the diagonal of the C-LIM matrix: Matrix Reasoning in the upper left cell; Picture Concepts in the middle cell; the mean of Similarities, Vocabulary, and Comprehension in the lower right cell. Participants were labeled as having scores that represented invalid estimates of their ability if the C-LIM cell means followed the hypothesized pattern of highest score in the top left cell, lower score in the middle cell, and lowest score in the bottom right cell. Alternatively, participants were categorized as having scores that represented valid estimates of their ability if the C-LIM cell means followed any other pattern (see Figure 1).

Flanagan et al. (2007, 2013) did not provide a specific numerical value for the minimum discrepancy required between cells for a score pattern to be designated different; rather, they indicated that importance lies in “the relationships between the scores and the degree to which they form a pattern that is either consistent or inconsistent with the pattern of performance predicted by the matrix“ (Flanagan et al., 2007, p. 181). This suggests that a 1-point difference between cell means down the diagonal of the C-LIM might suffice to designate scores as invalid and does not take standard error of measurement into consideration. Therefore, a sensitivity analysis was conducted to ensure that the magnitude of score differences down the diagonal did not produce markedly discrepant results. In this analysis, the score difference between the top left, middle, and bottom right cells was assessed at all possible score differences from 0 to 26 points.

For both analyses, frequencies of the WISC-IV C-LIM decisions (i.e., valid vs. invalid) were compared to the true state of participants’ cultural and linguistic diversity (WISC-IV normative sample membership considered the standard and ELL sample membership deemed to be culturally/linguistically different) to compute diagnostic utility statistics (see McFall & Treat, 1999, for more information on diagnostic utility statistics). In the context of this study, sensitivity is the proportion of ELL participants displaying the invalid profile and specificity is the proportion of normative group participants exhibiting the valid profile. In contrast, positive predictive power is the conditional probability of membership in the ELL group, given a C-LIM invalid score profile and negative predictive power is the conditional probability of membership in the WISC-IV normative group, given a C-LIM valid score profile.

Because these diagnostic statistics are influenced by prevalence rates, these data were also plotted on a receiver operating characteristic curve (ROC) graph and the area under the curve (AUC) was used to quantify the ROC (Centor & Schwartz, 1985; Hanley & McNeil, 1982). ROCs are not influenced by prevalence rates (Swets, 1988) and the AUC is a universal metric that is interpreted as the probability that diagnostic test results from a randomly selected individual who is positive for a condition (i.e., different cultural and linguistic experiences) will be higher than a randomly selected individual who does not have the condition (i.e., standard cultural and linguistic experiences) across every possible C-LIM cut score (Metz, 1978; Streiner & Cairney, 2007; Swets, 1988). The binary AUC computation described in Cantor and Kattan (2000) was applied to estimate the AUC for the single C-LIM score. Finally, a nonparametric approach was used to fit the curve 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). The following threshold values have been suggested to aid the interpretation of the AUC: values between 0.50 and 0.70 characterize low accuracy, values between 0.70 and 0.90 characterize medium accuracy, and values between 0.90 and 1.00 characterize high accuracy (Streiner & Cairney, 2007; Swets, 1988). The diagonal line in the ROC graph represents chance accuracy and the accuracy of a decision increases as the ROC curve approaches the upper left-hand corner of the graph.

Given that children in the ELL sample spoke Spanish at home, by definition they were culturally and linguistically different from the WISC-IV normative sample (Wechsler, 2003b). Thus, the ELL sample was compared to the WISC-IV normative sample with the expectation that all of the ELL participants would display the invalid C-LIM profile of decline and that scores from all of the

Table 1
Means and Standard Deviations of the WISC-IV Subtest, Index, and FSIQ Scores for 86 ELL Students Tested for Special Education Eligibility and the WISC-IV Standardization Sample (N = 2,033)

	WISC-IV Sample		ELL		
WISC-IV score	*M*	*SD*	*M*	*SD*	*d*
BD	10.3	2.8	8.8	2.7	−0.54
SI	10.3	2.8	7.5	2.3	−1.01
DS	10.3	2.8	7.3	2.1	−1.08
PCn	10.3	2.8	9.5	2.7	−0.29
CD	10.3	2.8	8.8	2.9	−0.54
VC	10.3	2.7	6.9	2.1	−1.27
LN	10.3	2.8	7.6	2.7	−0.97
MR	10.3	2.8	9.1	2.4	−0.43
CO	10.3	2.7	8.4	2.3	−0.71
SS	10.3	2.7	9.0	2.7	−0.48
VCI	101.2	13.4	86.2	10.2	−1.13
PRI	101.8	13.3	95.0	11.1	−0.51
WMI	101.1	13.3	85.3	11.4	−1.19
PSI	101.8	13.5	94.0	13.1	−0.58
FSIQ	102.2	13.0	87.3	9.3	−1.16

Note. WISC-IV = Wechsler Intelligence Scales for Children—Fourth Edition; ELL = English language learner; 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 = Symbol Search; VCI = Verbal Comprehension Index; PRI= Perceptual Reasoning Index; WMI = Working Memory Index; PSI = Processing Speed Index; FSIQ = Full Scale IQ Score. Differences significant at $p < .001$ for all 15 score comparisons across WISC-IV and ELL groups.

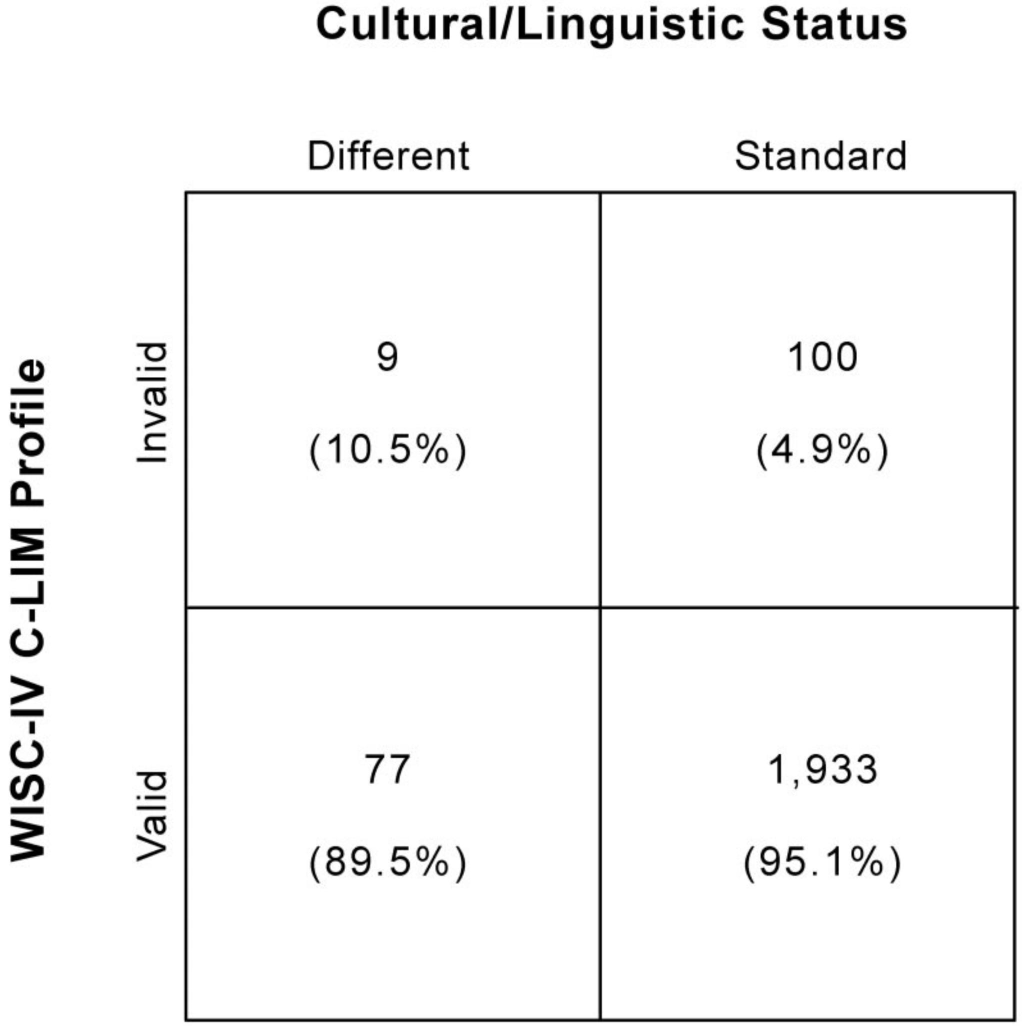
WISC-IV normative sample would be determined valid by the C-LIM. Although some ELL students with high levels of English proficiency might not display the invalid C-LIM pattern (Ortiz, 2011), they would be unlikely to substantially attenuate the overall diagnostic utility statistics. Thus, medium or high AUC scores should be obtained if the C-LIM is accurate with ROC curves near the upper right-hand corner of the graph.

Results

WISC-IV subtest, index, and FSIQ scores from the ELL sample were slightly lower and more varied than scores from the WISC-IV normative sample. Table 1 contains the means and standard deviations of all WISC-IV scores for participants disaggregated by participant subgroup. 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. Standardized mean differences ranged from −0.29 to −1.19. Mean subtest, index, and FSIQ scores for the referred sample of ELL students and the WISC-IV standardization sample are included in Table 1. The valid C-LIM profile (i.e., cell means did not decline) emerged in the mean WISC-IV subtest scores from both the WISC-IV normative sample and the ELL sample. Thus, neither sample of children exhibited the invalid C-LIM profile when group mean scores were considered.

In contrast, individual decisions based on C-LIM criteria are displayed in Figure 2. The true positive rate was 0.105, the true negative rate was 0.951, the false positive rate was 0.049, and the false negative rate was 0.895. Positive predictive power (conditional probability of membership in the ELL group, given a C-LIM invalid score profile) was 0.083 and the negative predictive power (conditional probability of membership in the WISC-IV normative group, given a C-LIM valid score profile) was 0.962.

To remove the effects of prevalence, sensitivity and 1-specificity were plotted on a ROC graph to investigate the probability that a randomly selected participant would be accurately classified by the C-LIM. An AUC value of 0.53 resulted when a binary ROC was used to compare C-LIM decisions between ELL and WISC-IV normative groups (see Figure 3). To illustrate ROC's resilience to prevalence rates, identical results were obtained when ten samples of $n = 86$ were randomly selected from the WISC-IV group and compared to the ELL group (*M* AUC = 0.53). AUC values ranged from 0.50 to 0.51 when all



		Cultural/Linguistic Status	
		Different	Standard
WISC-IV C-LIM Profile	Invalid	9 (10.5%)	100 (4.9%)
	Valid	77 (89.5%)	1,933 (95.1%)

Figure 2. Valid and invalid WISC-IV C-LIM profiles for 86 culturally and linguistically different (i.e., ELL sample) participants and 2,033 culturally and linguistically standard (i.e., WISC-IV normative sample) participants.

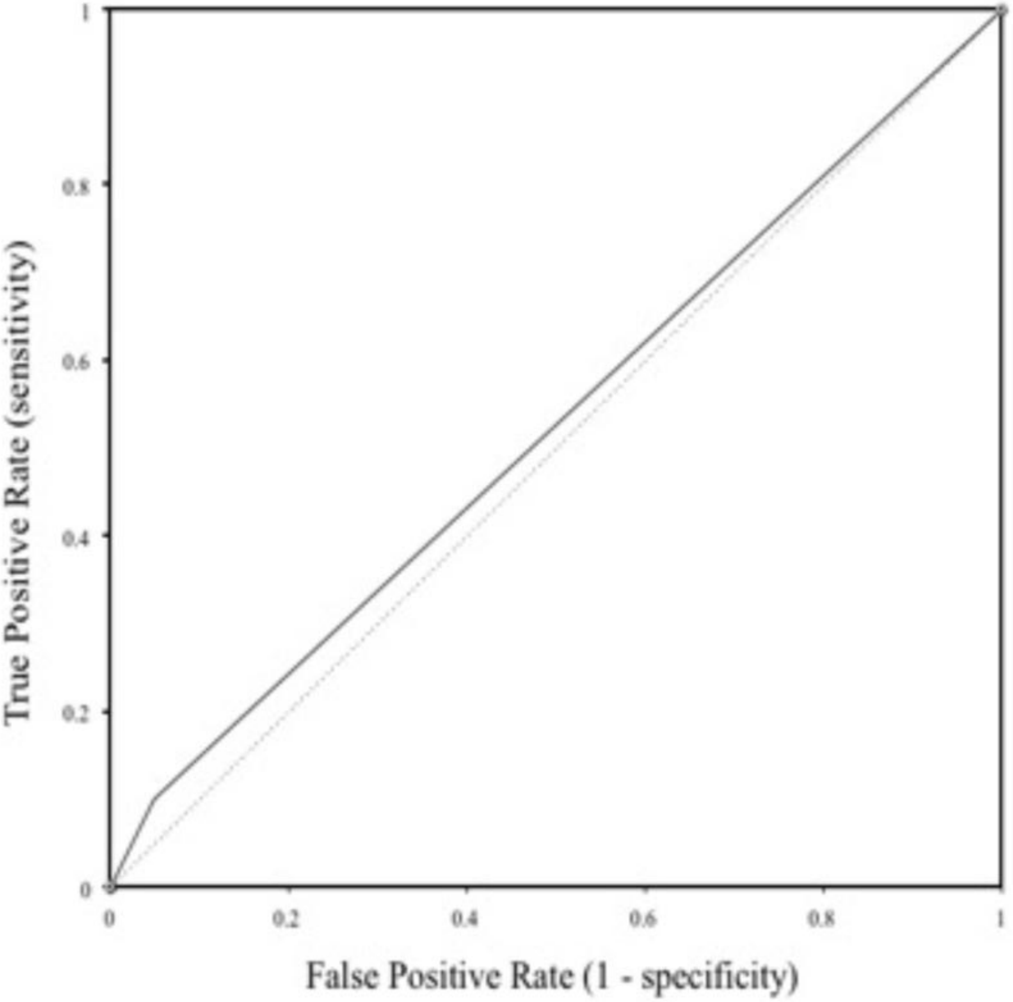


Figure 3. ROC graph illustrating the comparisons of true-positive and false-positive rates from a referred sample of ELL students (*n* = 86) and the WISC-IV normative sample (*n* = 2,033) when a single cut score was used.

possible cut scores were considered (see Figure 4). Thus, the probability that a randomly selected ELL participant would have a higher C-LIM score than a randomly selected normative group participant fell at chance levels (Streiner & Cairney, 2007; Swets, 1988).

Discussion

The purpose of the present study was to determine the degree to which the WISC-IV C-LIM can accurately distinguish culturally/linguistically different (i.e., ELL) children from culturally/linguistically standard (i.e., WISC-IV normative sample) children. As per Flanagan et al. (2007, 2013) and Ortiz (2011), ELL students should exhibit a profile of declining WISC-IV scores down the diagonal of the C-LIM (indicating invalid estimates of ability), whereas WISC-IV normative sample participants should exhibit some other profile of scores (indicating valid estimates of ability). On average, WISC-IV scores from the ELL children were significantly lower and more varied than scores from the WISC-IV

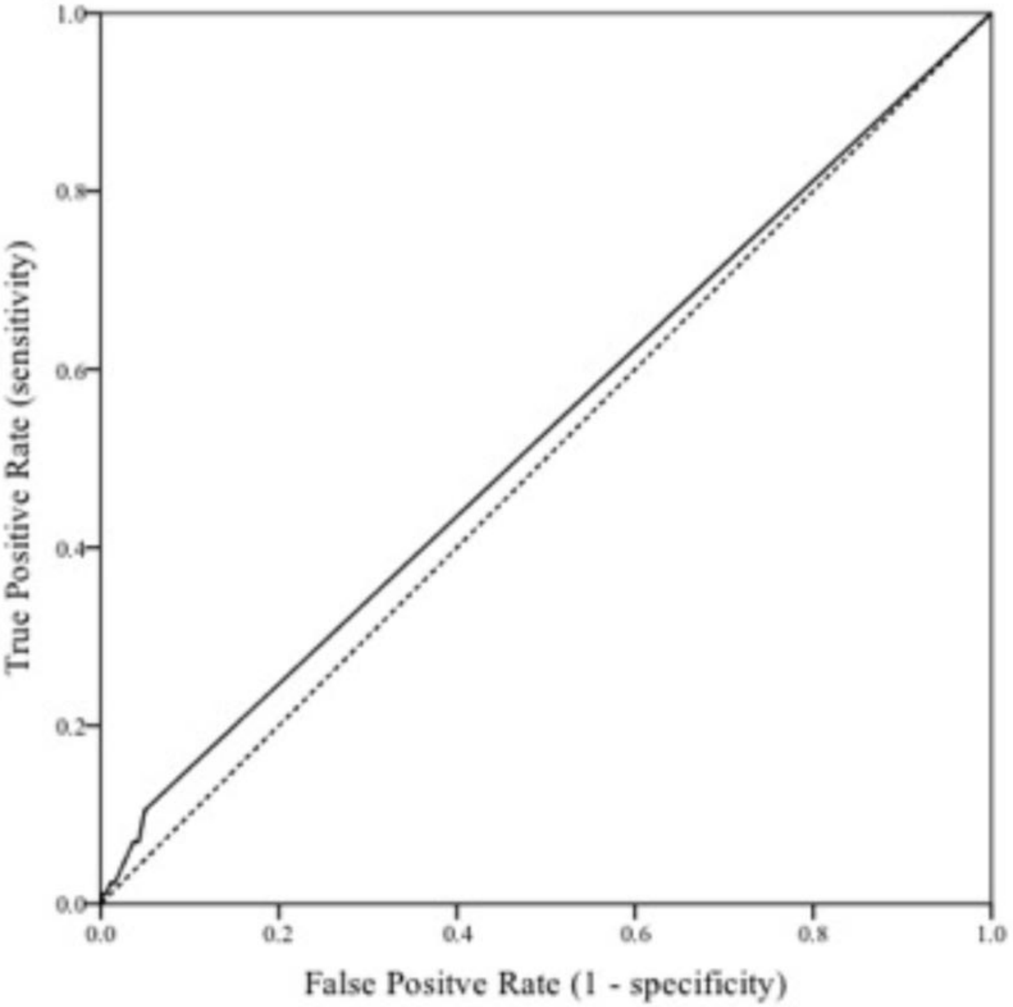


Figure 4. ROC graph illustrating the comparisons of true-positive and false-positive rates from a referred sample of ELL students (*n* = 86) and the WISC-IV normative sample (*n* = 2,033) when all possible cut scores were used.

normative sample. However, the invalid C-LIM profile was exhibited by only 10.5% of the ELL children. In contrast, scores for 4.9% of the children in the WISC-IV normative group were identified as invalid by the C-LIM. Scores from 100% of the ELL children and 0% of the normative group children should have been identified as invalid according to Flanagan et al. (2007, 2013) and Ortiz's (2011) hypotheses. Probabilistically, a randomly selected child from the ELL group and a randomly selected child from the norm group would be correctly identified by the C-LIM only 53% of the time.

Although there were mean differences in WISC-IV scores between the ELL and norm groups, this outcome has little relevance because significant mean score differences: (a) are not uncommon between referred samples and nonreferred samples, regardless of linguistic or cultural characteristics (Hale, 2010; Watkins, 2010); (b) have not been reliably identified across studies investigating the validity of the C-LIM (Kranzler et al., 2010; Tychanska, 2009; Verderosa, 2007); (c) are not indicative of test bias (Brown, Reynolds, & Whitaker, 1999; Reynolds, 2000); (d) do not signify clinical significance nor justify recommendations for clinical application to individual clients (Kraemer & Kupfer, 2006); and (e) constitute an inappropriate nomothetic approach to individual decision making.

Distributional overlap in scores between groups is so well known (Meehl & Rosen, 1955) that Meehl (1990) judged the failure to report an appropriate overlap measure as “unscholarly” (p. 232). Thus, decisions about individuals are more appropriately addressed with a diagnostic utility approach utilizing such statistics as the AUC (Kraemer et al., 2011; Kraemer & Kupfer, 2006; Swets et al., 2000). No previous study of the C-LIM used this approach. Kranzler et al. (2010) were the only researchers to report individual frequencies for the C-LIM, and they found that only 37% of individual participants from their sample of students receiving English as a second language services exhibited the hypothesized declining profile on the WJ III Cog.

Limitations

As with all applied research, the present study is not without limitations. First, the referred sample size was small and small sample sizes tend to inflate Type I and Type II error. The degree to which sample size influenced the results can be estimated by conducting a power analysis (Cohen, 1992). Power analyses require consideration of the statistical test conducted, effect size you expect to obtain, alpha level, and power level. The nonparametric AUC is mathematically equivalent to the Wilcoxon Mann-Whitney test (Hanley & McNeil, 1982) and Rice and Harris (2005) have tabulated the relationship between the AUC and Cohen’s *d* for AUC values ranging from 0.50 (i.e., chance diagnostic accuracy) to .97 (i.e., high diagnostic accuracy) when a 50% base rate is implied. However, there is no previous research investigating the C-LIM using ROC analyses from which to extract an AUC value to estimate an expected effect and Flanagan et al. (2007) did not specify a base rate for the percent of ELLs one would expect to follow the declining pattern. Nevertheless, only 50 participants per group would be necessary to detect a true effect for a test with medium accuracy and only 10 participants would be necessary per group to detect a true effect for a test with high accuracy assuming a 50% base rate (Rice & Harris, 2005; Streiner & Cairney, 2007; Swets, 1988).

A second limitation includes the absence of matching criteria on characteristics that may have influenced score variation such as age, gender, FSIQ score, and so on. WISC-IV IQ scores are based upon age normative groups (Wechsler, 2003b). 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 among participant subgroups may have affected scores. Gender is another variable that might affect 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 these differences are estimated to be small (Janet, 1981). Variation between participant groups on FSIQ scores may have also affected score variation as a result of Spearman’s law of diminishing returns, which suggests that the general cognitive ability factor (e.g., approximated by the 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; Der & Deary, 2003).

Lastly, the ELL group was composed of students referred for evaluation whose home language was Spanish according to parental report on a home language survey. No objective measure of language proficiency or cultural adaptation was included for these students. Consequently, some ELL students may have been more linguistically and culturally proficient than anticipated, and there was no way to identify how much their WISC-IV scores should have been affected by that variability. This possibility cannot be empirically evaluated, but AUC values near chance should only have been obtained if most or all of the ELL students were fluent in English. Although possible, this is not likely. Nevertheless, test scores from the ELL participants were extracted from archival files and it was not possible to control or manipulate conditions to better isolate the effects of culture and language to allow causal conclusions (Reinhart, Haring, Levin, Patall, & Robinson, 2013).

Conclusions

Future research should continue to address the clinical utility of the C-LIM. The present study only investigated the ability of the C-LIM to produce accurate decisions with WISC-IV scores, but there are 19 other standardized IQ tests for which Flanagan et al. (2007) have created C-LIMs that have yet to be empirically investigated. Furthermore, there are a variety of ways in which the C-LIM can be manipulated, such as by rearranging the subtest classifications or emphasizing the cell means at the extreme ends of the C-LIM (i.e., low linguistic demand/low cultural demand and high linguistic demand/high cultural demand). Nevertheless, it is essential that researchers use appropriate diagnostic utility methods for determining the clinical utility of the C-LIM before it is applied in clinical practice, especially given the results of the present study together with the cautionary conclusions from Kranzler et al. (2010) and Cormier (2012).

Flanagan et al. (2007) claimed that the C-LIM could reduce bias in test selection and interpretation, but this claim remains unsubstantiated. To the contrary, our results indicated that the C-LIM invalid profile was exhibited by only 10.5% of the ELL students. The lack of discriminatory power associated with the C-LIM profile calls into question the validity of using test scores from individuals to separate cultural and linguistic differences from SLD with diverse students.

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 postulated that profile analysis continues to be advanced in spite of unsubstantiated evidence because of intuitive appeal and an overreliance on clinical judgment (Garb, 2005). More recently, 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).

The C-LIM has been advocated in a number of books and workshops (e.g., www.crossbattery.com; Flanagan et al., 2007, 2013; Ortiz, 2011), and its developers have claimed that it is “based on the application of both prior and current empirical research” (Flanagan et al., 2013, p. 343). Yet, no empirical studies have been published in peer-reviewed journals to validate use of this tool. In contrast, inconclusive results have been reported in multiple dissertations regarding the reliability and validity of C-LIM decisions (Cormier, 2012; Tychanska, 2009; Verderosa, 2007) and a single peer-reviewed article conveyed results that did not substantiate its use in applied practice (Kranzler et al., 2010). Given that the burden of proof in science rests with the individual making a claim (Lilienfeld, Lynn, & Lohr, 2003), it is the responsibility of advocates of the C-LIM to implement a systematic research program (Levin & O’Donnell, 1999) designed to produce scientifically credible evidence sufficient to support use of the C-LIM for such high-stakes decisions as diagnosis of SLD (Marley & Levin, 2011).

Ethical codes require that psychologists use assessment instruments with established validity and reliability for the population being tested, and for the purpose of the assessment (American Psychological Association Ethics Code, 2002; National Association of School Psychologists, 2010; American Educational

Research Association, American Psychological Association, & National Council for Measurement in Education, 1999). Clearly, ethical and professional standards place responsibility on users of assessment methods (Gould, Martindale, & Flens, 2013). The C-LIM has intuitive appeal and is supported by engaging theoretical treatises, scientifically undocumented assertions of expert opinion, unwarranted generalizations from group differences, appeals to authority, case studies, and assurances of ethical probity (Gambrill, 2012). However, the model lacks empirical evidence of reliability and validity, and should not be employed until that evidential lacuna has been bridged (Baron, 1994; Cromer, 1993; Frisby, 2013; Lilienfeld, Ammirati, & David, 2012; Meehl, 1990; Stanovich, 2010).

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, & 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.

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.

Aziz, N. (2011). *Patterns of cognitive performance for culturally and linguistically diverse individuals with global cognitive impairment* (Doctoral dissertation). Retrieved from ProQuest (Accession No. 3441046).

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.

Baron, J. (1994). *Thinking and deciding* (2nd ed.). New York, NY: Cambridge University Press.

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

Brown, J. E. (2008). *The use and interpretation of the Batería III with U.S. bilinguals* (Doctoral dissertation). Retrieved from ProQuest (Accession No. 3343757).

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

Cantor, S. B., & Kattan, M. W. (2000). Determining the area under the ROC curve for a binary diagnostic test. *Medical Decision Making, 20,* 468–470.

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. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

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.

Cormier, D. C. (2012). *The influences of linguistic demand and cultural loading on cognitive test scores* (Doctoral dissertation). Retrieved from ProQuest (Accession No. 3513323).

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

Cromer, A. (1993). *Uncommon sense: The heretical nature of science.* New York, NY: Oxford University Press.

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.

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

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

Elliot, C. D. (1990). *Differential Ability Scales: Administration manual.* San Antonio, TX: The Psychological Corporation.

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

Fan, X., Willson, V. L., & Kapes, J. T. (1996). Ethnic group representation in test construction samples and test bias: The standardization fallacy revisited. *Educational and Psychological Measurement, 56,* 365–382.

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: Wiley.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). Cross-battery assessment of individuals from culturally and linguistically diverse backgrounds. In A. S. Kaufman & N. L. Kaufman (Series Ed.), *Essentials of cross-battery assessment* (3rd ed., pp. 287–350). Hoboken, NJ: Wiley.

Frisby, C. L. (2013). Testing, assessment, and cultural variation: Challenges in evaluating knowledge claims. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.), *Oxford handbook of child psychological assess-*

ment (pp. 150–171). New York, NY: Oxford University Press.

Gambrill, E. (2012). *Propaganda in the helping professions.* New York, NY: Oxford University Press.

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology, 1,* 67–89. doi:10.1146/annurev.clinpsy. 1.102803.143810

Gould, J. W., Martindale, D. A., & Flens, J. R. (2013). In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.), *Oxford handbook of child psychological assessment* (pp. 222–235). New York, NY: Oxford University Press.

Hale, C. (2010). *A cluster analytic study of the WISC-IV in children referred for psychoeducational assessment due to persistent academic difficulties.* University of Windsor, Windsor, Ontario, Canada.

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.

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 w2 and d. *American Psychologist, 36,* 892–901. doi:10.1037/0003-066x .36.8.892

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

Jordan, N. C., Glutting, J., Ramineni, C., & Watkins, M. W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review, 39,* 181–195.

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

Kraemer, H. C., Frank, E., & Kupfer, D. J. (2011). How to assess the clinical impact of treatments on patients, rather than the statistical impact of treatments on measures. *International Journal of Methods in Psychiatric Research, 20,* 63–72. doi: 10.1002/mpr. 340

Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry, 59,* 990–996. doi: 10.1016/j.biopsych. 2005.09.014

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.

Levin, J. R., & O'Donnell, A. M. (1999). What to do about educational research's credibility gaps? *Issues in Education: Contributions from Educational Psychology, 5,* 177–229.

Lilienfeld, S. O., Ammirati, R., & David, M. (2012). Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology, 50,* 7–36. doi:10.1016/j.jsp.2011.09.006

Lilienfeld, S. O., Lynn, S. J., & Lohr, J. M. (2003). *Science and pseudoscience in clinical psychology.* New York NY: Guilford.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40. doi:10.1037//1082–989X. 7.1.19

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

Marley, S. C., & Levin, J. R. (2011). When are prescriptive statements in educational research justified? *Educational Psychology Review, 23,* 197–206. doi: 10.1007/s10648-01-1-9154-y

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.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215–241.

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

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244.

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

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.

Metz, C. E. (2008). ROC analysis in medical imaging: A tutorial review of the literature. *Radiological Physics & Technology, 1,* 2–12. doi:10.1007/12194-007-0002-1

National Association of School Psychologists. (2010). *Principles for professional ethics.* Bethesda, MD: Author.

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).

Ortiz, S. O. (2011). Separating cultural and linguistic differences (CLD) from specific learning disability (SLD) in the evaluation of diverse students: Difference or disorder? In D. P. Flanagan & V. C. Alfonso (Eds.), *Essentials of specific learning disability identification* (pp. 299–325). Hoboken, NJ: Wiley.

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.

Reinhart, A. L., Haring, S. H., Levin, J. R., Patall, E. A., & Robinson, D. H. (2013). Models of not-so-good behavior: Yet another way to squeeze causality and recommendations for practice out of correlational data. *Journal of Educational Psychology, 105,* 241–247. doi:10.1037/a0030368

Reynolds, C. R. (2000). Why is psychometric research on bias in mental testing so often ignored? *Psychology,* *Public Policy, and Law, 6*(1), 144–150. doi: 10.1037//1076 -8971.6.1.144

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

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

Roid, G. H. (2003). *Stanford-Binet intelligence scales* (5th ed.). Itasca, IL: Riverside Publishing.

Sandoval, J., & Miille, M.P.W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology, 48,* 249–253.

Souravlis, S. A. L. (2010). *Evaluating speech-language and cognitive impairment patterns via the Culture-Language Interpretive Matrix* (Doctoral dissertation). Retrieved from ProQuest (Accession No. 3435601).

Stanovich, K. E. (2010). *How to think straight about psychology* (9th ed.). Boston, MA: Pearson.

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristics curves. *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, DC.

Verderosa, F. A. (2007). *Examining the effects of language and culture on the Differential Ability Scales with bilingual preschoolers* (Doctoral dissertation). Retrieved from ProQuest (Accession No. 3286027).

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.

Weiner, I. B. (2003). Prediction and postdiction in clinical decision making. *Clinical Psychology: Science and Practice, 10,* 335–338. doi:10.1037//1082-989X. 7.1.19

Weiss, D. J., Shanteau, J., & Harries, P. (2006). People who judge people. *Journal of Behavioral Decision Making, 19,* 441–454. doi:10.1002/bdm.52d

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.

Wiggins, J. S. (1987). *Personality and prediction: Principles of personality assessment.* Malabar, FL: Krieger Publishing.

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

Woodcock, R. W., Muñoz-Sandoval, A. F., McGrew, K. S., & Mather, N. (2007). *Batería-III.* Itasca, IL: Riverside Publishing.

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

Date Received: September 29, 2012
Date Accepted: September 23, 2013
Associate Editor: Michael Vanderwood ■

Kara. M. Styck is an assistant professor in the School Psychology program at The University of Texas at San Antonio. Her research interests include the detection and prevention of errors in high-stakes decisions made by school psychologists and the use of advanced quantitative methods to evaluate and develop empirically based classification systems.

Marley W. Watkins is Professor and Chairman of the Department of Educational Psychology at Baylor University. Dr. Watkins is a Diplomate of the American Board of Professional Psychology and a member of the Society for the Study of School Psychology. His research interests include professional issues, the psychometrics of assessment and diagnosis, individual differences, and computer applications.