Contemp School Psychol
DOI 10.1007/s40688-014-0021-y

18, 168-177.

Discriminant Validity of the WISC-IV Culture-Language Interpretive Matrix

Kara M. Styck · Marley W. Watkins

Abstract The Culture-Language Interpretive Matrix (C-LIM) was developed to help practitioners determine the validity of test scores obtained from students who are culturally and linguistically different from the normative group of a test. The present study used an idiographic approach to investigate the diagnostic utility of the C-LIM for the Wechsler Intelligence Scales for Children-Fourth Edition (WISC-IV) for distinguishing between a referred sample of 69 English language learners, 79 English-speakers diagnosed with Autism, and 216 English speakers referred for a special education evaluation. Results indicated that the WISC-IV C-LIM differentiated between these groups of students at chance rates. Evidence from the previous studies along with the results of the current study does not support the use of the C-LIM for making decisions about individuals in applied practice.

Keywords C-LIM · Diagnostic utility · ROC · WISC-IV · English Language Learners

Discriminant Validity of the WISC-IV Culture-Language Interpretive Matrix

Standardized IQ tests play an integral role in the high-stakes decisions made by school psychologists regarding eligibility for special programming (Reschly and Hosp 2004). IQ test scores have norm-referenced interpretations, which mean that the performance of an examinee is described relative to the performance of participants in the normative sample. This has led some researchers to believe that IQ test scores may be invalid when an examinee does not share cultural and linguistic experiences with a sufficient proportion of the normative group of a test (Flanagan and Ortiz 2001; Flanagan et al. 2007, 2013; Oriz and Dynda 2005; Oriz et al. 2012; Ortiz 2011; Ortiz and Ochoa 2005).

Consequently, Flanagan et al. (2013) encouraged the use of the Culture-Language Interpretive Matrix (C-LIM) "to assist [practitioners] in determining whether results obtained from standardized testing are valid (and may therefore be interpreted) or not" (p. 310). The C-LIM is a 3×3 matrix originated by Flanagan and Ortiz (2001) with rows and columns that represent increasing linguistic and cultural demand, respectively. Subtests from individual standardized IQ tests are placed into the cells of test-specific C-LIMs according to their hypothesized cultural and linguistic demand classifications (Flanagan and Ortiz 2001). Figure 1 illustrates a C-LIM for the Wechsler Intelligence Scales for Children-Fourth Edition (WISC-IV; Wechsler 2003a), which is one of the most popularly used IQ tests by school psychologists (Braden and Athanasiou 2013).

During the formation of the C-LIM, Flanagan and Ortiz relied upon group mean score differences between English language learners (ELL) and English speakers, as well as an expert consensus procedure to classify subtests from various IQ tests as having low, moderate, or high cultural and linguistic demand. However, the research cited by the C-LIM authors as providing empirical evidence in support of the culture-language test classification system, "included mean scores for bilingual individuals on various tests, most commonly the Wechsler batteries" (Flanagan et al. 2007, p. 169). This suggests that clinical judgment was the primary method used to classify subtests as having low, moderate, or high cultural and linguistic demand, which is known to be less reliable than

K. M. Styck (✉)
Department of Educational Psychology, University of Texas at San Antonio, 501 W. Cesar E. Chavez Blvd., San Antonio, TX 78207-4415, USA
e-mail: Kara.Styck@utsa.edu

M. W. Watkins
Department of Educational Psychology, Baylor University, Waco, TX, USA

Published online: 21 June 2014

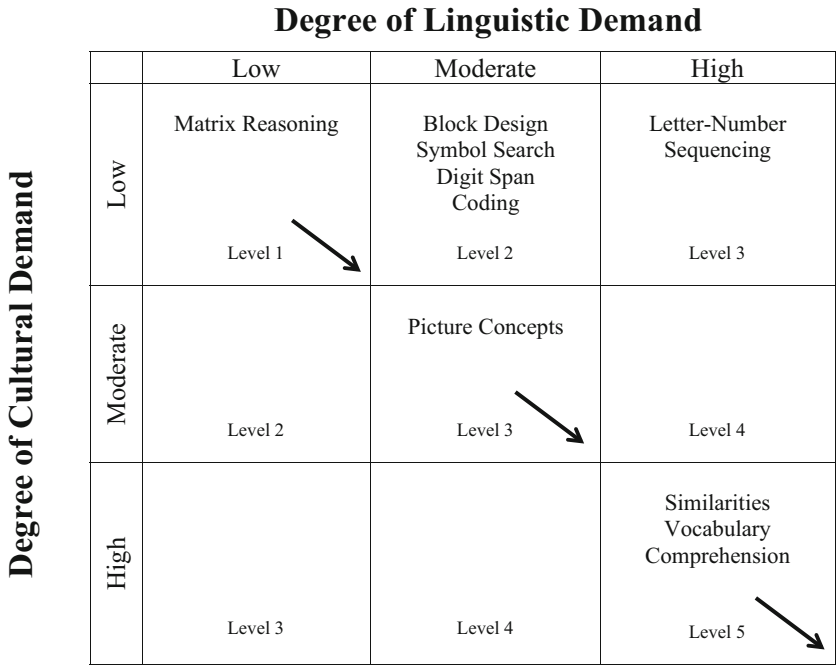

Fig. 1 A Culture-Language Interpretive Matrix for the Wechsler Intelligence Scale for Children-Fourth Edition. *Note.* Levels 1–5 represent the lowest (Level 1) to highest (Level 5) degree of score attenuation expected due to linguistic and cultural demand, and the *black arrows* represent the profile of decline hypothesized to indicate test scores are invalid (Flanagan et al. 2013)

actuarial methods (Dawes and Corrigan 1974; Dawes et al. 1989). Furthermore, several important issues remain unclear regarding the expert consensus procedure described by Flanagan et al. (2007, 2013), such as the qualifications of each expert, the number of experts involved, the degree to which intra- or inter-rater reliability estimates were computed to estimate the reliability of subtest classifications within and between experts, and the manner in which expert disagreements were resolved.

According to Flanagan et al. (2013), test scores are invalid if the cell means of the C-LIM systematically decline down the diagonal. If any other pattern emerges, results of standardized testing are considered valid (Flanagan and Ortiz 2001; Flanagan et al. 2007, 2013; Ortiz 2011, 2013; Ortiz and Ochoa 2005). Furthermore, in the presence of the invalid profile, Flanagan et al. (2007) asserted that "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" (p. 197). This strongly suggests that C-LIM decisions are interpreted independent of other information obtained throughout the comprehensive evaluation process. There is no known base rate for the percent of culturally and linguistically diverse students that can be expected to follow the profile of decline and the evidence offered to justify the interpretation of the C-LIM profiles consists mostly of unpublished doctoral dissertations—none of which reported the frequency of individual study participants who exhibited each C-LIM profile despite its purported use to make decisions about individuals (Aziz 2011; Dhaniram-Beharry 2008; Durandisse 2013; Souravlis 2010; Nieves-Brull 2006; Sotello-Dynega 2007; Sotello-Dynega et al. 2013; Templeton 2012; Tychanska 2009; Verderosa 2007).

The nomothetic analytical approach used in the majority of research on the C-LIM does not adequately address the degree to which the C-LIM can make accurate decisions for individual children and adolescents suspected of having a disability (Kraemer et al. 2011; Weiner 2003). Nomothetic analytical approaches investigate group differences, whereas idiographic analytical approaches investigate individual differences. Evaluating the accuracy of the C-LIM for making individual decisions requires the use of an idiographic analytical approach, such as the computation of sensitivity and specificity statistics. Sensitivity describes the proportion of ELL test scores that exhibit the invalid profile (i.e., decline in cell means down the diagonal) and specificity describes the proportion of test scores from English-speaking students that follow the valid profile (i.e., no decline in cell means down the diagonal). Kranzler et al. (2010) and Styck and Watkins (2013) are the only studies to date on the C-LIM published in peer-reviewed journals that reported diagnostic utility statistics. Kranzler et al. (2010) indicated

that only 37 % of students enrolled in English as a second language programming followed the C-LIM profile of decline on the Woodcock–Johnson Tests of Cognitive Abilities-Third Edition (WJ-III COG; Woodcock et al. 2001) C-LIM and Styck and Watkins (2013) reported that merely 11 % of a sample of ELLs referred to determine eligibility for special education programming followed the C-LIM profile of decline on the WISC-IV C-LIM.

Recently, two unpublished dissertations described similar negative results. Meyer (2013) and Van Deth (2013) reported sensitivity ranging between 0 and 35 % and specificity ranging between 84 and 100 % when the C-LIM invalid profile was used to compare test scores from the WJ-III COG and the Kaufman Assessment Battery for Children-Second Edition (KABC-II; Kaufman and Kaufman 2004) between samples of ELLs and English speakers referred for special education evaluations using a variety of criteria to define a systematic decline in cell means down the diagonal of the matrix. Altogether, empirical research on the C-LIM using an idiographic approach suggests that C-LIM decisions have low sensitivity and high specificity. Given these results, the C-LIM has not demonstrated the ability to accurately identify students who are culturally and linguistically different from those who share similar cultural and linguistic experiences with test normative groups.

Nevertheless, a diagnostic sign is useful if it can distinguish between clinical and non-clinical subgroups, referred subgroups with and without a particular condition, and if it can aid in making differential diagnoses. Previous research on the C-LIM using an idiographic approach has focused on the former two types of judgments by comparing a referred sample of ELLs with non-referred English speakers for the WISC-IV and WJ-III Cog C-LIMs (Meyer 2013; Styck and Watkins 2013) and by comparing a referred sample of ELLs with English speakers for the KABC-II C-LIM (Van Deth 2013). However, the ability of the C-LIM to distinguish between a referred sample of ELLs and English speakers identified as having a disability marked with significant impairments in cognitive ability or those with language impairments that might also attenuate scores as the degree of linguistic demand increases (i.e., Intellectual Disability, Autism, or Speech and Language Impairment, American Psychiatric Association 2013; Individuals with Disabilities Education Improvement Act 2004), remains unknown.

In spite of the absence of this critical information, Flanagan et al. (2013) claimed that, "Practitioners may rest assured that to date, no other factor has been discovered, apart from cultural and linguistic difference, that results in or has the capacity to create a declining pattern of test performance" (p. 323). Therefore, the purpose of the present study is to empirically test this assertion. Test scores from an English speaker diagnosed with Autism may decline as the linguistic demand of subtests increase because individuals with Autism are characterized by significant deficits in social communication skills (American Psychiatric Association 2013). However, their scores would *not* be expected to follow the invalid profile of decline indicating that the combined influence of cultural and linguistic differences attenuated their performance *nor* would their scores be expected to decline as the cultural demand of subtests increases according to Flanagan et al. (2007, 2013) hypotheses. Therefore, the presence of either of these two profiles of decline (i.e., due to the combined effect of culture and language or due to the separate effect of culture) in the test scores of students identified with Autism would suggest that the C-LIM is not able to distinguish between individuals whose cultural experiences differ.

Furthermore, previous research on the diagnostic utility of the C-LIM has compared non-referred samples of English speakers to referred samples of ELLs (Meyer 2013; Styck and Watkins 2013; Van Deth 2013). However, referred samples have different distributional characteristics than non-referred samples (Canivez and Watkins 1998; Chen and Zhu 2012; Watkins and Smith 2013), and these results may not generalize to applied situations in which test scores are only examined by practitioners for individuals who were referred to them for evaluations. As a result, a secondary purpose of the present investigation is to determine the degree to which the C-LIM can distinguish between a referred sample of ELLs and a referred sample of English speakers. It is hypothesized that the C-LIM will not be able to distinguish between either of these two groups of students because of its failure to materialize in meaningful numbers for any sample to date (Kranzler et al. 2010; Meyer 2013; Styck and Watkins 2013; Van Deth 2013).

Method

Participants

Participants included 364 school-aged children and adolescents who were referred for a multidisciplinary evaluation to determine eligibility as students in need of special education services in two Southwestern school districts and subsequently identified by a school multidisciplinary evaluation team as (a) having limited English proficiency according to the results of a home language survey, (b) having Autism, or (c) ineligible for special education services. The ELL sample included 69 students (46 males, 23 females) aged 6 to 16 years (M= 11.4, SD=2.8) whose native and home languages were both reported as Spanish. All ELL received special education services, with approximately 89 % identified as having specific learning disabilities and the remaining 11 % identified as having a variety of disabilities including emotional

disturbance, attention-deficit hyperactivity disorder, hearing impairment, other health impairment, and speech and language impairment.

The sample of English speakers identified as having autism included 79 students (62 males, 17 females) aged 6 to 16 years (M=10.4, SD=2.7). IDEA (2004) includes other developmental disabilities under the eligibility category of Autism, such as Asperger's disorder and Rett's disorder. Approximately 95 % of the students in the Autism sample received a primary diagnosis of Autism and 5 % of the students in the Autism sample received a primary diagnosis of Asperger's disorder. Finally, the sample of English speakers identified as ineligible for special education services included 216 (140 males, 76 females) students aged 6 to 16 years (M=9.7, SD=2.2).

Instruments

The WISC-IV is an individually administered and standardized intelligence test for children and adolescents. It is composed of a standard battery of ten core subtests (M=10; SD=3) that form four index composite scores, the Verbal Comprehension Index (VCI), Working Memory Index (WMI), Processing Speed Index (PSI), Perceptual Reasoning Index (PRI), and a Full-Scale IQ score (FSIQ; M=100; SD=15). The WISC-IV standardization sample included 2,200 children ages 6 years and 0 months to 16 years and 11 months who were fluent in English and represented the 2,000 United States census stratified on age, sex, race, ethnicity, parent education level, and geographic region (Wechsler 2003b). Test developers and independent researchers have provided evidence of reliability and validity (Chen and Zhu 2008; Watkins et al. 2006; Wechsler 2003b).

Procedure

WISC-IV scores, parent home language survey results, and disability eligibility status were collected from archival special education records from two large Southwestern school districts following university institutional review board and school district approval. No other identifying information was collected from participant files.

Students were included in the ELL sample if (a) a parent home language survey indicated that Spanish was their native language as well as the primary language spoken at home, (b) they were not identified as having Autism by a school multidisciplinary evaluation team, and (c) their file contained subtest scores for the WISC-IV core battery. Students were included in the Autism sample if (a) a parent home language survey indicated that English was their native language as well as the only language spoken at home and (b) their file contained subtest scores for the WISC-IV core battery. Finally, a sample of students was included in the study if (a) a parent home language survey indicated that English was their native language as well as the only language spoken at home and (b) they were found ineligible for special education services by a school multidisciplinary evaluation team.

Analyses

All analyses were conducted using the base package of R Version 3.0.1 (R Core Team 2013). First, a one-way analysis of variance (ANOVA) was computed to investigate the degree to which mean WISC-IV subtest scaled scores, composite scores, and FSIQ scores significantly differed between participant subgroups. The Welch approximate F test was used to evaluate the omnibus test and post hoc pairwise comparisons to relax the assumption of homogeneity of variances. Finally, the Bonferonni correction was applied to maintain an experimentwise error rate of 0.05.

Next, sensitivity and specificity statistics were computed according to the presence or absence of the profile of decline (Flanagan et al. 2007, 2013; Ortiz and Ochoa 2005). Frequencies of the WISC-IV C-LIM patterns (i.e., valid versus invalid) were compared to the true state of participants' cultural and linguistic difference (McFall and Treat 1999). According to Flanagan et al. (2013) hypotheses, ELL participants should exhibit the invalid profile (decline of scores in the diagonal of the C-LIM) and English-speaking participants with Autism or those referred for special education evaluations should exhibit the valid profile (any pattern of scores other than decline of scores in the diagonal of the C-LIM). Therefore, sensitivity is defined as the proportion of ELL participants whose WISC-IV scores follow the invalid profile, and specificity is defined as the proportion of participants diagnosed with Autism or found ineligible for special education services whose WISC-IV scores follow the valid profile.

In addition, the hypothesized separate influences of linguistic demand and cultural demand were evaluated by inspecting the degree to which scores systematically declined for each participant group as linguistic and cultural demand increased. Scores from ELL students and English speaking students with Autism should decline across the columns of the C-LIM as linguistic demand increases, but scores from English-speaking students with Autism should not decline across the rows of the C-LIM as cultural demand increases according to Flanagan et al. (2013). Finally, scores from a referred sample of English-speaking students should not systematically decline as a function of either the hypothesized increasing linguistic or cultural demand according to Flanagan et al. (2013).

Sensitivity and specificity statistics provide useful information about the accuracy of decisions made on the basis of a single cut-score. However, plotting 1-specificity and sensitivity on the respective x- and y-axes of a receiver operating characteristic (ROC; Cantor and Kattan 2000) graph can provide a more useful index of diagnostic accuracy. Diagnostic accuracy improves as the sensitivity increases and the false

positive rate (i.e., 1-specificity) decreases. This is illustrated by an x coordinate near the origin and a y coordinate near the top of the graph. The area under the curve (AUC) quantifies this information and is interpreted as the probability that the WISC-IV scores from a randomly selected ELL participant will follow the invalid profile of decline and the WISC-IV scores from a randomly selected English-speaking participant diagnosed with Autism or a randomly selected English-speaking participant referred for a special education evaluation will exhibit a valid profile (Centor and Schwartz 1985; Hanley and McNeil 1982). Therefore, chance accuracy in distinguishing between valid and invalid WISC-IV subtest score patterns is represented by a diagonal line. AUC values ranging between 0.50 and 0.70 represent low accuracy, whereas AUC values between 0.70 and 0.90 characterize medium accuracy, and AUC values between 0.90 and 1.00 indicate high accuracy (Streiner and Cairney 2007; Swets 1988). Medium or high accuracy should be attained if the WISC-IV C-LIM can distinguish between ELLs and English speakers with Autism or a referred sample of English speakers.

The AUC value obtained from a non-parametric ROC graph is mathematically identical to the Wilcoxon Mann–Whitney test when outcome data are continuous (Hanley and McNeil 1982). This suggests that approximately 50 participants per group would be required to accurately identify a test with medium accuracy, and approximately 10 participants per group would be required to accurately identify a test with high accuracy (Faul et al. 2007). However, Flanagan et al. (2013) did not specify the base rate of students who are culturally and linguistically different from the normative group of a test that can be expected to follow the invalid profile of decline and this value may not be 50 %. Furthermore, the present investigation evaluated the accuracy of the C-LIM as it was intended to be used (i.e., yes/no decision afforded by the presence/absence of the invalid profile of decline). Therefore, the estimated sample sizes designated above represent approximations derived from the information currently available.

Results

Mean WISC-IV scores were statistically different between groups for most subtest scaled scores, composite scores, and the FSIQ score: Similarities $F(2, 147.0)=27.7$, $p<0.001$; Digit Span $F(2, 136.6)=14.5$, $p<0.001$; Picture Concepts $F(2, 134.7)=8.2$, $p<0.001$; Coding $F(2, 131.7)=10.1$, $p<0.001$; Vocabulary $F(2, 135.8)=34.5$, $p<0.001$; Letter-Number Sequencing $F(2, 132.3)=19.5$, $p<0.001$; Comprehension $F(2, 119.0)=12.1$, $p<0.001$; Symbol Search $F(2, 140.2)=15.1$, $p<0.001$; VCI $F(2, 127.4)=24.7$, $p<0.001$; PRI $F(2, 138.3)=11.5$, $p<0.001$; WMI $F(2, 128.9)=22.9$, $p<0.001$; PSI $F(2, 134.6)=15.7$, $p<0.001$; FSIQ $F(2, 134.7)=29.3$, $p<0.001$. However, post hoc pairwise comparisons indicated that mean WISC-IV scores from the ELL group were statistically lower than both English-speaking groups only for the Similarities ($d=-1.05$ vs. Autism group, $d=-0.90$ vs. ineligible group), Digit Span ($d=-0.68$ vs. Autism group, $d=-0.72$ vs. ineligible group), Picture Concepts ($d=-0.06$ vs. Autism group, $d=-0.50$ vs. ineligible group), and Vocabulary ($d=-0.89$ vs. Autism group, $d=-1.12$ vs. ineligible group) subtests as well as the VCI ($d=-0.67$ vs. Autism group, $d=-1.01$ vs. ineligible group) and WMI ($d=-0.72$ vs. Autism group, $d=-0.97$ vs. ineligible group). Table 1 contains mean WISC-IV subtest scores, composite scores, and FSIQ scores disaggregated by participant group with significant differences indicated at the $p<0.05$, 0.01, and 0.001 levels after adjusting for multiple statistical tests using the Bonferonni correction.

Results for the comparison of ELL students and students with Autism indicated the WISC-IV C-LIM had a true positive rate of 4.3 % and a true negative rate of 88.6 %. This resulted in a false positive rate of 11.4 % (i.e., $(1–0.886)\times 100=11.4$ %), which is larger than the true positive rate of 4.3 %. The probability was 46.5 % that WISC-IV scores from a randomly selected participant in the ELL group would be identified as invalid by the C-LIM and the WISC-IV scores from a randomly selected participant in the group of students identified as having Autism would be identified as valid by the C-LIM. This AUC value represents low diagnostic accuracy (Streiner and Cairney 2007; Swets 1988).

Similar results were obtained when the C-LIM was used to distinguish between ELLs and a referred sample of English-speaking students: the WISC-IV C-LIM had a true positive rate of 21.4 % and a true negative rate of 75.6 %. This yields a false positive rate of 24.4 % (i.e., $(1–0.756)\times 100=24.4$ %), which is larger than the true positive rate of 21.4 %. The probability was 48.5 % that the WISC-IV scores from a randomly selected participant in the ELL group would follow the pattern of decline down the diagonal of the C-LIM and WISC-IV scores from a randomly selected participant in the group of English-speaking students referred for special education evaluations would not follow the pattern of decline down the diagonal of the C-LIM. This AUC value represents near chance accuracy (Streiner and Cairney 2007; Swets 1988). Figure 2 contains a graph illustrating the ROC curves for each comparison when the accuracy of the C-LIM profile of decline was evaluated (i.e., combined influence of cultural and linguistic demand).

Inspection of the accuracy of the C-LIM profile of decline across the increasing influence of linguistic demand and cultural demand, separately, yielded comparable results (Figs. 3 and 4). Approximately 32.9 % of scores from English-speaking students with Autism declined as the hypothesized linguistic demand of subtests increased, but 15.2 % of scores from this group also declined as the hypothesized cultural demand of

Table 1 Means and standard deviations of the WISC-IV Subtest, Index, and FSIQ Scores for all Participants Disaggregated by Study Subgroup

	Autism (n=69)		ELL (n=79)		Ineligible (n=216)	
WISC-IV Score	*M*	*SD*	*M*	*SD*	*M*	*SD*
BD	9.9	3.7	8.6	2.6	9.8	2.9
SI	10.0$_{***}$	3.0	7.1$^{***}_{***}$	2.4	9.4***	2.7
DS	8.8**	3.5	6.7$^{**}_{***}$	2.7	8.6$_{***}$	2.6
PCn	9.0$_{**}$	3.4	8.8$^{***}_{**}$	2.7	10.1***	2.5
CD	7.0**	3.8	8.4	2.9	9.1**	2.6
VC	9.1$_{***}$	3.8	6.3$^{***}_{***}$	2.5	9.1***	2.5
LN	8.5	3.7	6.7***	2.9	9.2***	2.6
MR	9.8	3.5	8.6	2.6	9.8	2.6
CO	7.9	4.3	7.6**	3.2	9.4**	2.2
SS	7.6***	3.2	8.5	2.6	9.6***	2.6
VCI	94.4$_{**}$	19.7	83.1$^{***}_{**}$	13.1	95.5***	11.4
PRI	97.3	18.9	91.8***	11.6	99.7***	12.5
WMI	92.1$_{**}$	18.2	80.5$^{***}_{**}$	13.8	93.0***	11.8
PSI	84.9***	17.5	91.3	13.0	96.3***	12.4
FSIQ	91.0	18.3	83.4***	11.4	95.4***	11.5

Note. WISC-IV Wechsler Intelligence Scale 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, and *FSIQ* Full Scale IQ Score

$^{*}p<0.05$

$^{**}p<0.01$

$^{***}p<0.001$

subtests increased. Furthermore, approximately 22.7 % of scores from a sample of English-speaking students referred for a special education evaluation declined as the hypothesized linguistic demand of subtests increased and 11.1 % of scores from this participant group also declined as the cultural demand of subtests increased. Likewise, comparisons between groups when the accuracy of the separate hypothesized influence of linguistic and cultural demands were inspected produced AUC values ranging between 53.3 and 61.8 %, which fall within chance accuracy (Streiner and Cairney 2007; Swets 1988).

Discussion

ELL scores were lower than the WISC-IV normative sample. However, these are common characteristics of referred samples regardless of cultural and/or linguistic status (Canivez and Watkins 1998; Chen and Zhu 2012; Watkins and Smith 2013), which is evident in the descriptive statistics reported for the English-speaking samples in the present study as well. Statistically, significant mean score differences emerged between participant groups on some of the WISC-IV subtests, index scores, and the FSIQ score. However, results of the ROC analyses indicated that the WISC-IV C-LIM did not distinguish between a referred sample of ELLs and English-speaking students found ineligible for special education services or between a referred sample of ELLs and English-speaking students diagnosed with Autism. AUC values were at chance levels for all comparisons—even when the influence of cultural and linguistic demand was evaluated separately. Moreover, a randomly selected participant from both samples of English-speaking students had a slightly *higher* probability of following the C-LIM profile of decline (i.e., combined influence of cultural and linguistic demand) than a randomly selected participant from the sample of ELLs (Fig. 2).

Flanagan et al. (2013) commented that results of Kranzler et al. (2010) were likely due to the "limited sample size [rather] than with any inherent problems in the classifications" (p. 315) because cell differences down the diagonal of the C-LIM "would be only about three to four points apart at most" (p. 315). However, interpretation of such small differences ignores measurement error and effectively treats observed scores as error-free. Of additional concern, Flanagan (2014) responded to the results of Styck and Watkins (2013) by stating in an online discussion forum that "Identification of ELLs is not the purpose of the C-LIM… [when] performance systematically declines as a function of increasing culture and language test demands based on the C-LIM results, then the validity of the WISC-IV results is called into question." Following this logic, a test is invalid for ELL students who exhibit the invalid C-LIM test profile but valid for ELL students who exhibit the valid C-LIM test profile. Given this argument, how is the C-LIM decision, either valid or invalid, to be verified other than by referral to the C-LIM profile itself?

No tool yields decisions that are 100 % accurate. Scales are not perfectly calibrated, rulers are only nearly the same length, and scores obtained from psychological test instruments represent estimates of an individual's ability. The C-LIM is not an exception to this rule, and Flanagan et al. (2013) are encouraged to empirically evaluate the error rate of their tool. "The criterion of the scientific status of a theory is its falsifiability, or refutability, or testability" (Popper 2002, p. 48). The lack of falsifiability is a serious scientific flaw, along with emphasis on confirmation, lack of self-correction, evasion of peer review, and absence of boundary conditions (Lilienfeld et al. 2012). The present study included ELL students because it seemed reasonable to believe that they should display a greater proportion of C-LIM invalid profiles than other students if the C-LIM hypothesis was correct. We found that our ELL students could not be distinguished from other students on the basis of C-LIM profiles. Collectively, the accumulated evidence suggests that the C-LIM is inaccurate in making decisions about individuals (Kranzler et al. 2010; Styck and Watkins 2013; Meyer 2013; Van Deth 2013), which strongly

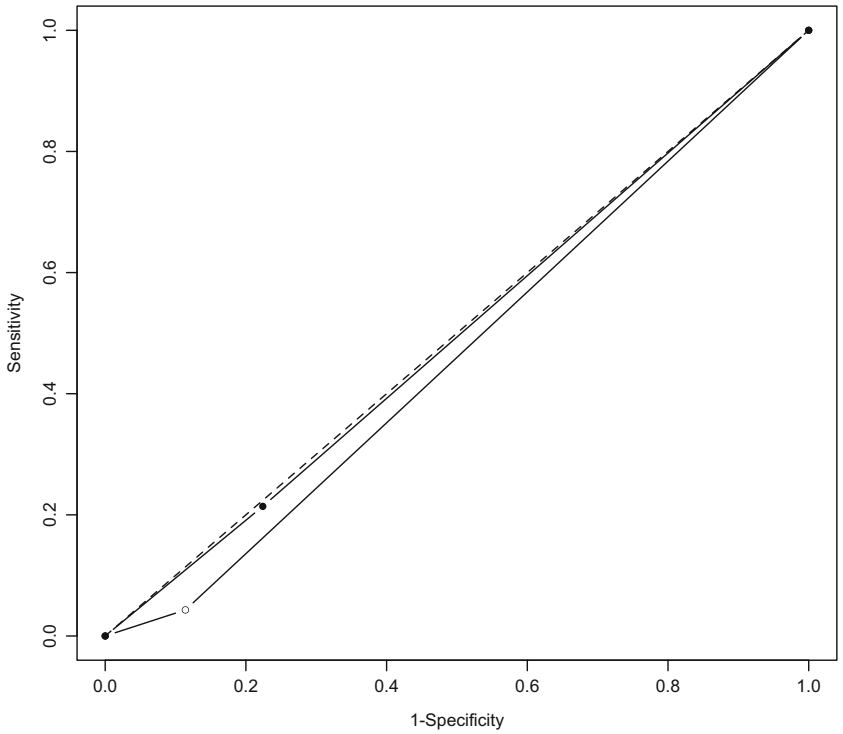

Fig. 2 ROC graph illustrating sensitivity and 1-specificity rates from a referred sample of ELLs (n=79) compared to a referred sample of English-speakers found ineligible for special education services by a multidisciplinary evaluation team (n=216) symbolized by the *black circle* and a referred sample of ELLs compared to a referred sample of English-speakers identified as having Autism (n=69) symbolized by the *white circle*. The *dotted-diagonal line* represents chance accuracy

suggests that it should not be used in applied practice. If this conclusion is disputed, it seems opportune to paraphrase Platt (1964) and ask what experiment could disprove the C-LIM hypothesis.

All studies contain limitations, and the present investigation is not an exception. First, the groups were formed based on the information obtained from archival special education records and it was not possible to empirically evaluate the English proficiency of the ELL group. Likewise, the Autism group was formed on the basis of school multidisciplinary evaluation team decisions assumed to be accurate. It is possible that the C-LIM profile of decline did not emerge in the

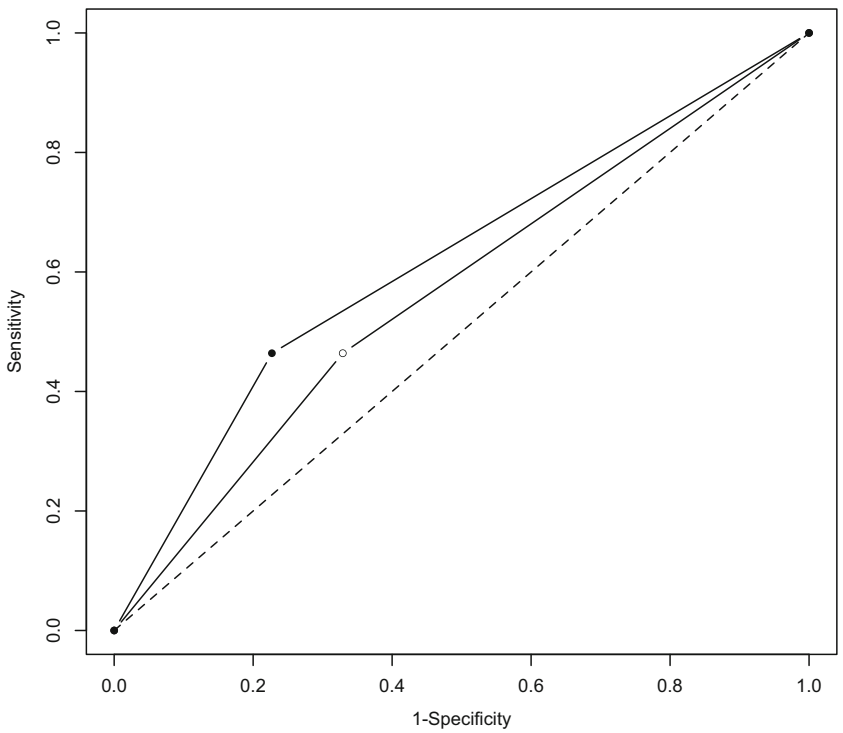

Fig. 3 ROC graph illustrating sensitivity and 1-specificity rates of linguistic demand from a referred sample of ELLs (n=79) compared to a referred sample of English-speakers found ineligible for special education services by a multidisciplinary evaluation team (n=216) symbolized by the *black circle* and a referred sample of ELLs compared to a referred sample of English-speakers identified as having Autism (n=69) symbolized by the *white circle*. The dotted-diagonal line represents chance accuracy

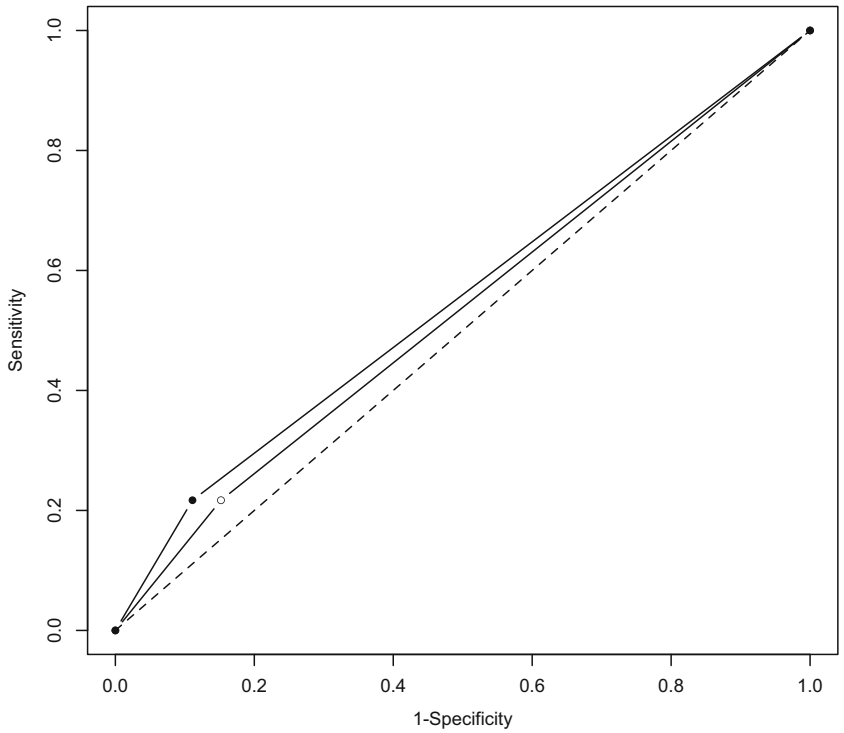

Fig. 4 ROC graph illustrating sensitivity and 1-specificity rates of cultural demand from a referred sample of ELLs (n=79) compared to a referred sample of English-speakers found ineligible for special education services by a multidisciplinary evaluation team (n=216) symbolized by the *black circle* and a referred sample of ELLs compared to a referred sample of English-speakers identified as having Autism (n=69) symbolized by the *white circle*. The *dotted-diagonal line* represents chance accuracy

sample of ELL students included in the present investigation because they were viewed by school psychologists as proficient in English and as sharing cultural experiences with the normative group of the test. However, the C-LIM is meant to be included as part of comprehensive individual psychoeducational evaluations. A comparison of scores from a non-referred group of ELLs to a non-referred group of English speakers would not generalize to this target population (Meehl and Rosen 1955). This is especially true given that referred samples tend to have lower means and standard deviations than non-referred samples (Canivez and Watkins 1998; Chen and Zhu 2012; Watkins and Smith 2013). Nevertheless, results of the present investigation have important implications for practitioners who work with culturally and linguistically diverse students.

Perhaps the most puzzling aspect of the C-LIM literature is that Flanagan et al. (2007, 2013) speak to differences in language and culture for individual children, while simultaneously encouraging practitioners to make decisions about individuals based upon group means. The message is "the average person in your group has X scores on this test, so *you* will also have X scores." But, individuals vary along many dimensions other than their degree of English language proficiency and acculturation to U.S. society, making the group norm a poor predictor of individual performance. Results of the present study illustrate this point. Significant mean differences were observed between participant subgroups on many WISC-IV scores, but C-LIM subtest score profiles distinguished between very few individuals. Proponents of the C-LIM may criticize the use of a single test battery for evaluating the C-LIM profiles in the present study due to the Flanagan et al. (2013) emphasis on cross-battery assessments. However, the examination of intra-individual score patterns (i.e., the degree to which an examinee's IQ subtest scores match the invalid score profile) is problematic for numerous reasons. These types of score interpretations have been widely denounced primarily due to the lack of diagnostic utility for score profiles (Devena and Watkins 2012; Smith and Watkins 2004; Watkins et al. 2002) and the temporal instability of subtest difference scores (Borsuk et al. 2006; McDermott et al. 1989a, b, 1992; Watkins and Smith 2013). Thus, the C-LIM was built on the shaky foundation of subtest score differences. Particularly, disquieting evidence has emerged from recent research on examiner bias (McDermott et al. 2013). Specifically, as much as 14 % of subtest variability could be attributed to examiners rather than examinees and each subtest was differentially vulnerable to examiner bias.

Given the importance of high-stakes decisions made by school psychologists for vulnerable populations, it is imperative that a rigorous program of research be conducted to significantly improve the accuracy of the C-LIM or, if that is not feasible, a more empirically sound approach to the assessment of culturally and linguistically diverse individuals be developed (Lilienfeld et al. 2003). Accordingly, the C-LIM is not recommended for use in applied settings until its extensive limitations are addressed and peer-reviewed research supports its use for making accurate decisions about individual students.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington: American Psychiatric Association.

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).

Borsuk, E. R., Watkins, M. W., & Canivez, G. L. (2006). Long-term stability of membership in a Wechsler Intelligence Scale for Children-Third Edition (WISC-III) subtest core profile taxonomy. *Journal of Psychoeducational Assessment, 24*, 52–68. doi:10.1177/0734282905285225.

Braden, J. P., & Athanasiou, M. S. (2013). Psychological assessment in school settings. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (2nd ed., Vol. 10, pp. 291–314). Hoboken: Wiley.

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

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.

Chen, H., & Zhu, J. (2008). Factor invariance between genders of the Wechsler Intelligence Scale for Children-Fourth Edition. *Personality and Individual Differences, 45*, 260–266. doi:10.1016/j.paid.2008.04.008.

Chen, H., & Zhu, J. (2012). Measurement invariance of WISC-IV across normative and clinical samples. *Personality and Individual Differences, 52*, 161–166. doi:10.1016/j.paid.2011.10.006.

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.

Devena, S. E., & Watkins, M. W. (2012). Diagnostic utility of WISC-IV General Abilities Index and Cognitive Proficiency Index difference scores among children with ADHD. *Journal of Applied School Psychology, 28*, 133–154. doi:10.1080/15377903.2012.669743.

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

Durandisse, M. (2013). *The effect of English language proficiency and acculturation on the Woodcock-Johnson Tests of Cognitive Abilities-Third Edition performance: In a Haitian-Creole sample* (Doctoral dissertation). Retrieved from ProQuest (Accession No. 3569999).

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Flanagan, D. P. (2014, January 15). Re: Testing EAL Students [Online forum comment]. Retrieved from http://school-psych-talk.freeforums.net/post/29/quote/27

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

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 (Eds.), *Essentials of cross-battery assessment second edition* (2nd ed., pp. 146–205). Hoboken: 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 (Eds.), *Essentials of cross-battery assessment* (3rd ed., pp. 287–350). Hoboken: Wiley.

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.

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

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.

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.

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

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.

McDermott, P. A., Glutting, J. J., Jones, J. N., & Noonan, J. V. (1989a). Typology and prevailing composition of core profiles in the WAIS-R standardization sample. *Psychological Assessment. A Journal of Consulting and Clinical Psychology, 1*, 118–125.

McDermott, P. A., Glutting, J. J., Jones, J. N., Watkins, M. W., & Kush, J. (1989b). Core profile types in the WISC-R national sample: Structure, membership, and applications. *Psychological Assessment. A Journal of Consulting and Clinical Psychology, 1*, 292–299.

McDermott, P. A., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *The Journal of Special Education, 25*, 504–526.

McDermott, P. A., Watkins, M. W., & Rhoad, A. M. (2013). Whose IQ is it? Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment, 26*, 207–214. doi:10.1037/a0034832.

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

Meyer, E. X. (2013). *Diagnostic accuracy of the Culture-Language Interpretive Matrix with the WJ-III NU: A comparison of Spanish speaking English language learners and monolingual English-speaking students* (Unpublished doctoral dissertation). The Pennsylvania State University, College Station, PA.

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).

Oriz, S. O., & Dynda, A. M. (2005). Use of intelligence tests with culturally and linguistically diverse populations. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 545–556). New York: Guilford.

Oriz, S. O., Ochoa, H. S., & Dynda, A. M. (2012). Testing with culturally and linguistically diverse populations: Moving beyond the verbal-performance dichotomy into evidence-based practice. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (3rd ed., pp. 526–552). New York: Guilford.

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.

Ortiz, S. O. (2013). *Sample interpretive statements for use with the Culture-Language Interpretive Matrix (C-LIM)*. Retrieved from http://www.crossbattery.com/

Ortiz, S. O., & Ochoa, S. H. (2005). Cognitive assessment of culturally and linguistically diverse individuals, an integrated approach. In R. L. Rhodes, S. H. Ochoa, & S. O. Ortiz (Eds.), *Assessing culturally and linguistically diverse students: a practical guide* (pp. 168–201). New York, NY: Guilford.

Platt, J. R. (1964). Strong inference. *Science, 146*, 347–353.

Popper, K. (2002). *Conjectures and refutations: the growth of scientific knowledge*. New York: Routledge.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: http://www.R-project.org/

Reschly, D. J., & Hosp, J. L. (2004). State SLD identification policies and practices. *Learning Disabilities Quarterly, 27*, 197–213.

Smith, C. B., & Watkins, M. W. (2004). Diagnostic utility of the Bannatyne WISC-IV pattern. *Learning Disabilities Practice, 19*, 46–56.

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

Sotello-Dynega, M., Ortiz, S. O., Flanagan, D. P., & Chaplin, W. F. (2013). English language proficiency and test performance: an evaluation of bilingual students with the Woodcock-Johnson III Tests of Cognitive abilities. *Psychology in the Schools, 50*, 781–797. doi:10.1002/pits.21706.

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).

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.

Styck, K. M., & Watkins, M. W. (2013). Diagnostic utility of the Culture-Language Interpretive Matrix for the WISC-IV among referred students. *School Psychology Review, 42*, 367–382.

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

Templeton, M. M. (2012). *An examination of the effects of culture and language on the executive functioning of Spanish-speaking English learners according to the Delis-Kaplan Executive Function System* (Doctoral dissertation). Retrieved from ProQuest (Accession No. 3505517).

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).

Van Deth, L. X. (2013). *Validity of the KABC-II Culture-Language Interpretive Matrix: A comparison of native English speakers and Spanish-speaking English language learners* (Unpublished doctoral dissertation). The Pennsylvania State University, College Station, PA.

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., & Smith, L. G. (2013). Long-term stability of the Wechsler Intelligence Scale for Children-Fourth edition. *Psychological Assessment, 25*, 477–483. doi:10.1037/a0031653.

Watkins, M. W., Kush, J. C., & Schaefer, B. A. (2002). Diagnostic utility of the learning disability index. *Journal of Learning Disabilities, 35*, 98–103.

Watkins, M. W., Wilson, S. M., Kotz, K. M., Carbone, M. C., & Babula, T. (2006). Factor structure of the Wechsler Intelligence Scale for Children-Fourth Edition among referred students. *Educational and Psychological Measurement, 66*, 976–983. doi:10.1177/0013164406288168.

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

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

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

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 development and evaluation of empirically-based classification systems, psychometrics, individual differences, and professional issues in school psychology.

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.