C-LIM Myths and Misconceptions

WHAT THE RESEARCH REALLYS SAYS.

Virtual Presentation for the
Multilingual Interest Group
April 21, 2023

Samuel O. Ortiz, Ph.D.
St. John’s University

The Culture-Language Interpretive Matrix (C-LIM)

An example of translation of research into practice for evaluating test score validity

The C-LIM is a systematic framework for organizing and guiding evidence-based practice. Its ONLY purpose is to provide a systematic method for evaluating the extent to which cultural and linguistic factors may have compromised test score validity. There are 3 basic criteria that provide evidence to suggest that test performance reflects the primary influence of cultural and linguistic factors and not actual ability, or lack thereof. These criteria are:

1. Overall pattern of decline from low language/culture tests to high language/culture tests.
2. All scores within or above expected range of "difference."
3. No significant score variability among scores within the same cells or between cells of the same level.

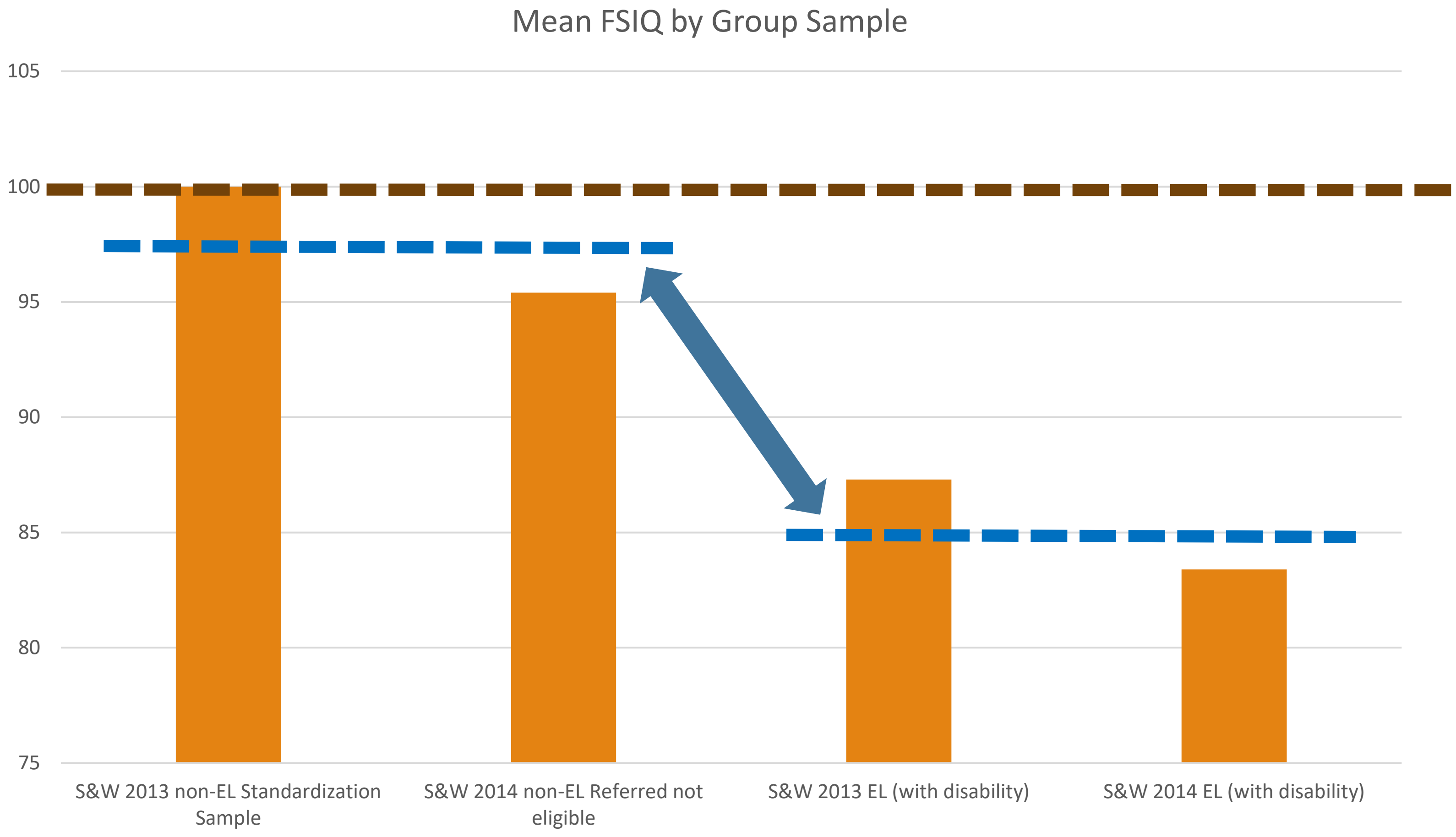
Results are **LIKELY INVALID** if **ALL** conditions are **MET.** Results are **LIKELY VALID** when **ANY** condition is **NOT MET.**

The C-LIM is NOT:

1. a diagnostic tool;
2. a test, scale, or attempt to measure anything other than the impact of culture and language;
3. an ML identification system;
4. a static framework or approach;
5. based on new research, rather, its underlying principles span more than a century.

Research Foundations for ML Evaluation

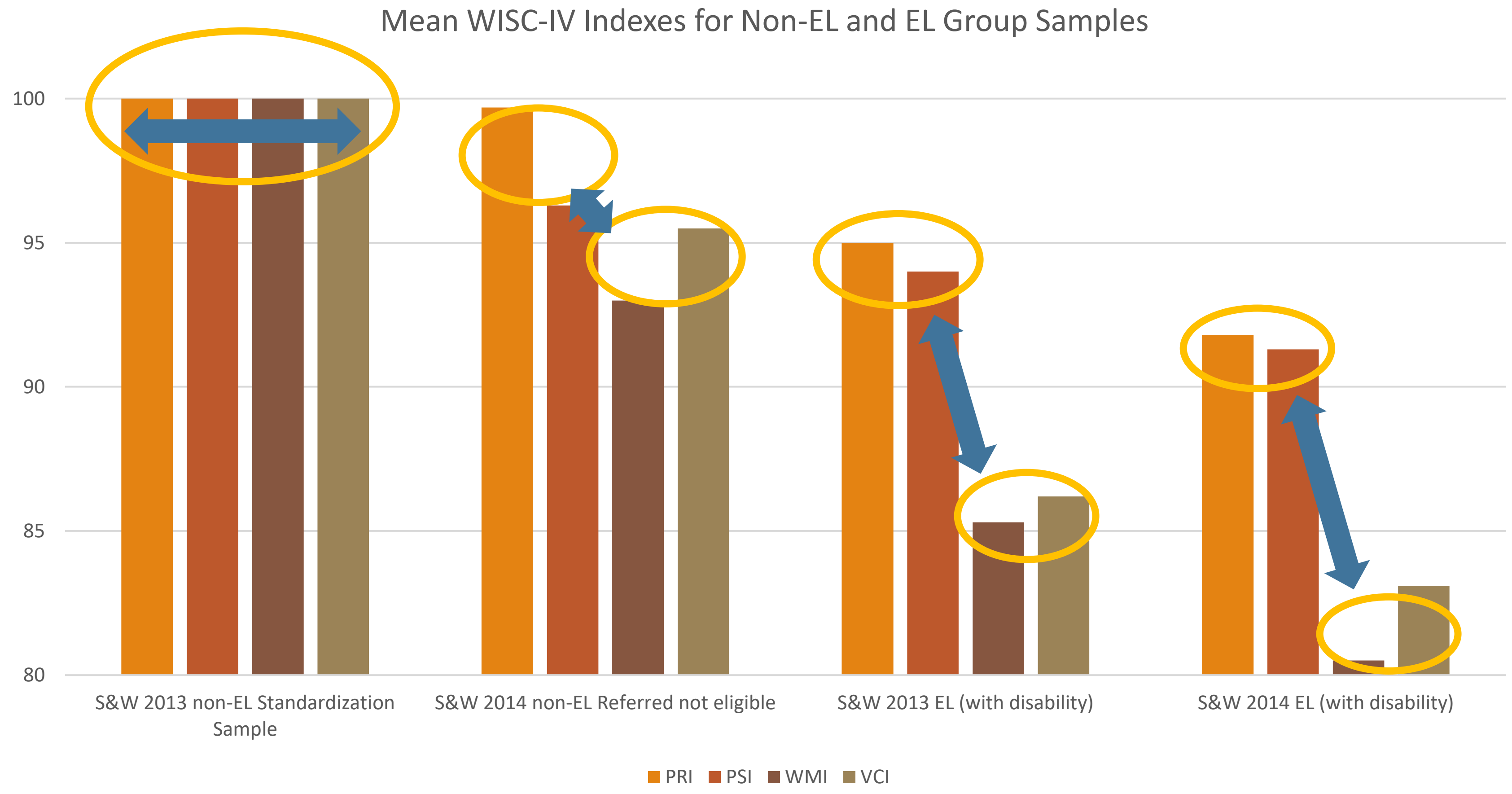
Research Principle 1: MLs and non-ML's perform differently at the broad ability level



*Source: Styck, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382.

Research Foundations for ML Evaluation

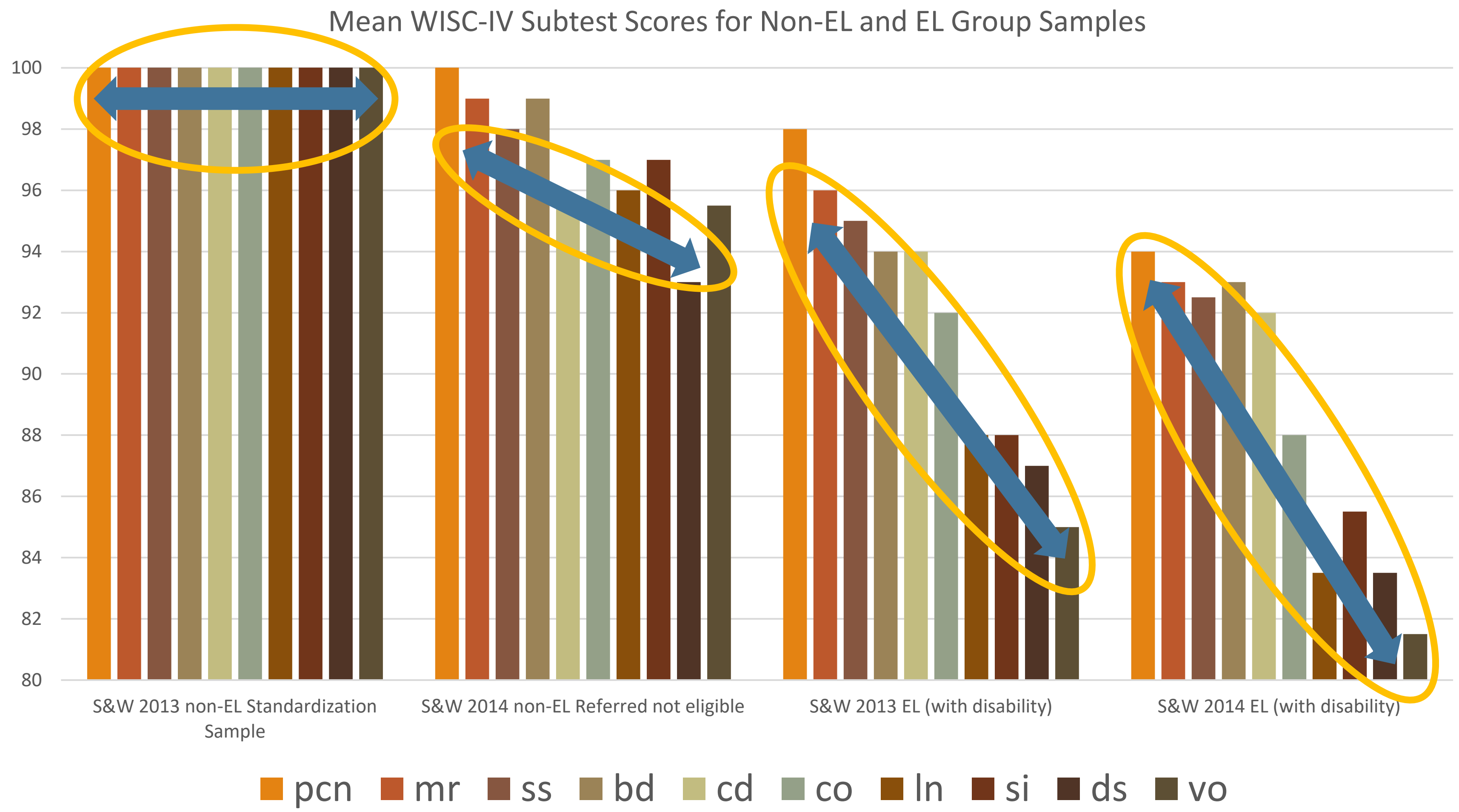
Research Principle 2: MLs perform better on nonverbal tests than verbal tests



*Source: Styck, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382.

Research Foundations for ML Evaluation

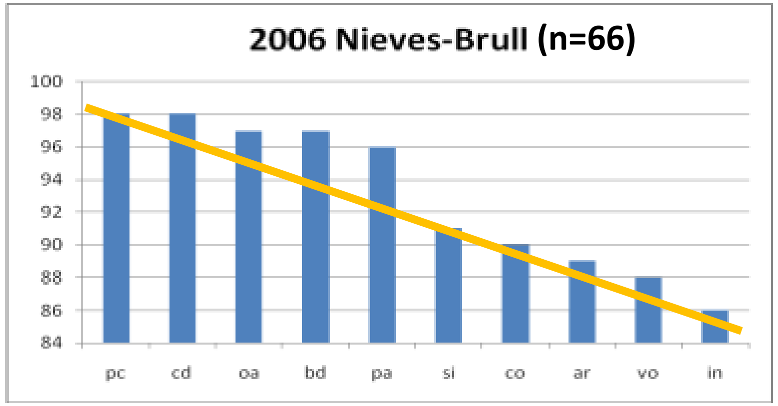
Research Principle 3: ML performance is moderated by linguistic/acculturative variables

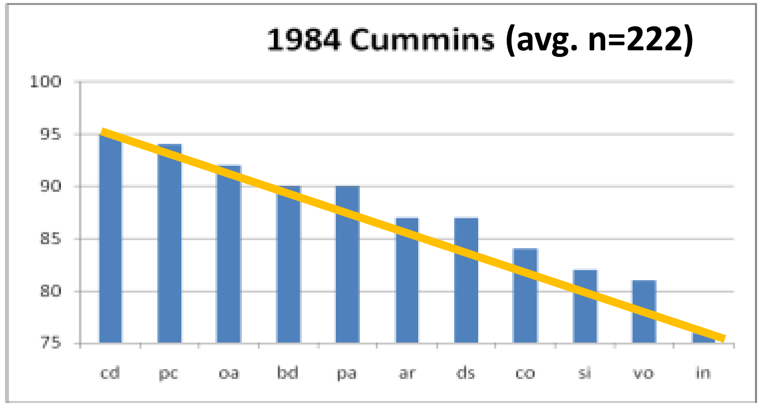


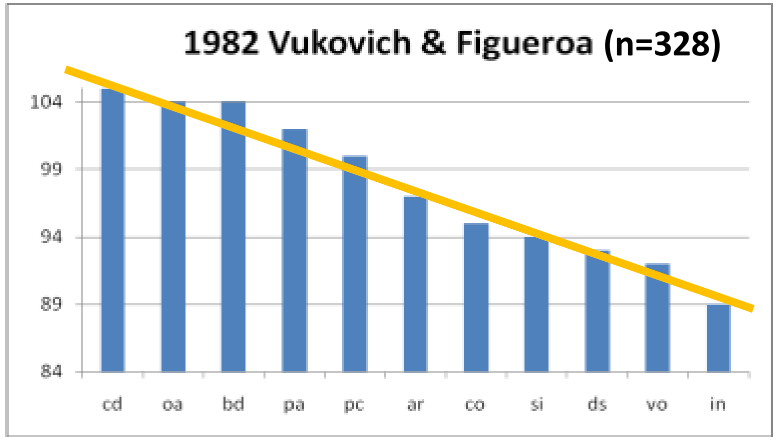
*Source: Styck, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382.

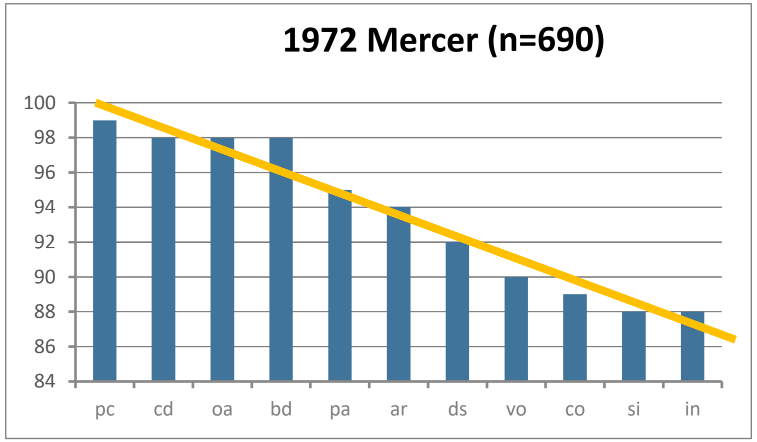
Research Foundations for ML Evaluation

Research Principle 3: ML performance is moderated by linguistic/acculturative variables









Research Foundations for ML Evaluation

Research Principle 3: ML performance is moderated by linguistic/acculturative variables

Highest Language Demands → Lowest Language Demands

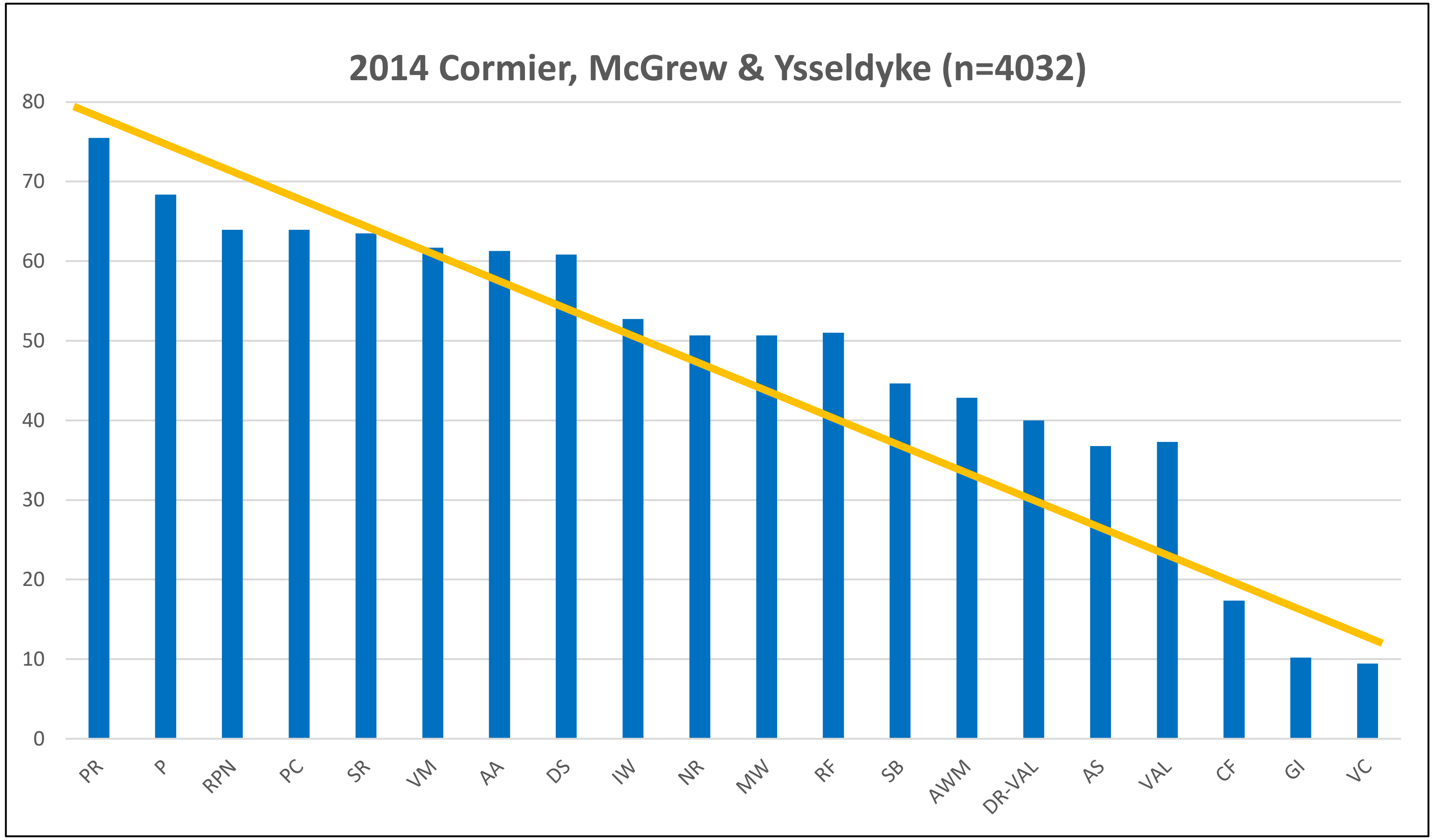
Table 3. Variance Explained by Exogenous Variables (Individual Test Performance) by Age Group.

Individual test	Variance explained 7-10	11-14	15-18	
Verbal Comprehension	.79[c]	.86[c]	.81[c]	1
General Information	.71[c]	.85[c]	.86[c]	
Concept Formation	.67[c]	.71[c]	.67[c]	2
Visual–Auditory Learning	.40[b]	.37[b]	.41[b]	
Delayed Recall Visual–Auditory Learning	.39[b]	.32[b]	.37[b]	
Analysis Synthesis	.29[b]	.44[b]	.47[b]	
Sound Blending	.25[b]	.32[b]	.35[b]	3
Auditory Working Memory	.22[b]	.44[b]	.32[b]	
Retrieval Fluency	.22[b]	.22[b]	.28[b]	
Memory for Words	.18[b]	.32[b]	.23[b]	
Numbers Reversed	.17[b]	.26[b]	.30[b]	
Pair Cancelation	.17[b]	.11[b]	.11[b]	
Rapid Picture Naming	.16[b]	.07[a]	.16[b]	
Incomplete Words	.13[b]	.31[b]	.23[b]	4
Visual Matching	.13[b]	.15[b]	.16[b]	
Decision Speed	.12[b]	.15[b]	.19[b]	
Auditory Attention	.10[b]	.20[b]	.15[b]	
Spatial Relations	.08[a]	.16[b]	.16[b]	5
Planning	.07[a]	.12[b]	.11[b]	
Picture Recall	.02[a]	.06[a]	.10[b]	

**Source: Cormier, D.C., McGrew, K.S. & Ysseldyke, J. E. (2014). The Influences of Linguistic Demand and Cultural Loading on Cognitive Test Scores. Journal of Psychoeducational Assessment, 32(7), 610-623.*

Research Foundations for ML Evaluation

Research Principle 3: ML performance is moderated by linguistic/acculturative variables



Research Foundations for ML Evaluation

Research Principle 3: ML performance is moderated by linguistic/acculturative variables

"the influence of language ability, *particularly receptive language ability*, is more influential than age on cognitive test performance. This last point highlights the importance of considering language abilities when assessing students' cognitive abilities." (p. 9) (Cormier et al., 2022)

Variable	β	R^2
Step 1		.66***
Age	.67***	
Step 2		.79***
Age	.37***	
Lifetime Exposure to English	.52***	
		ΔR^2 = .18***

*** p <.001

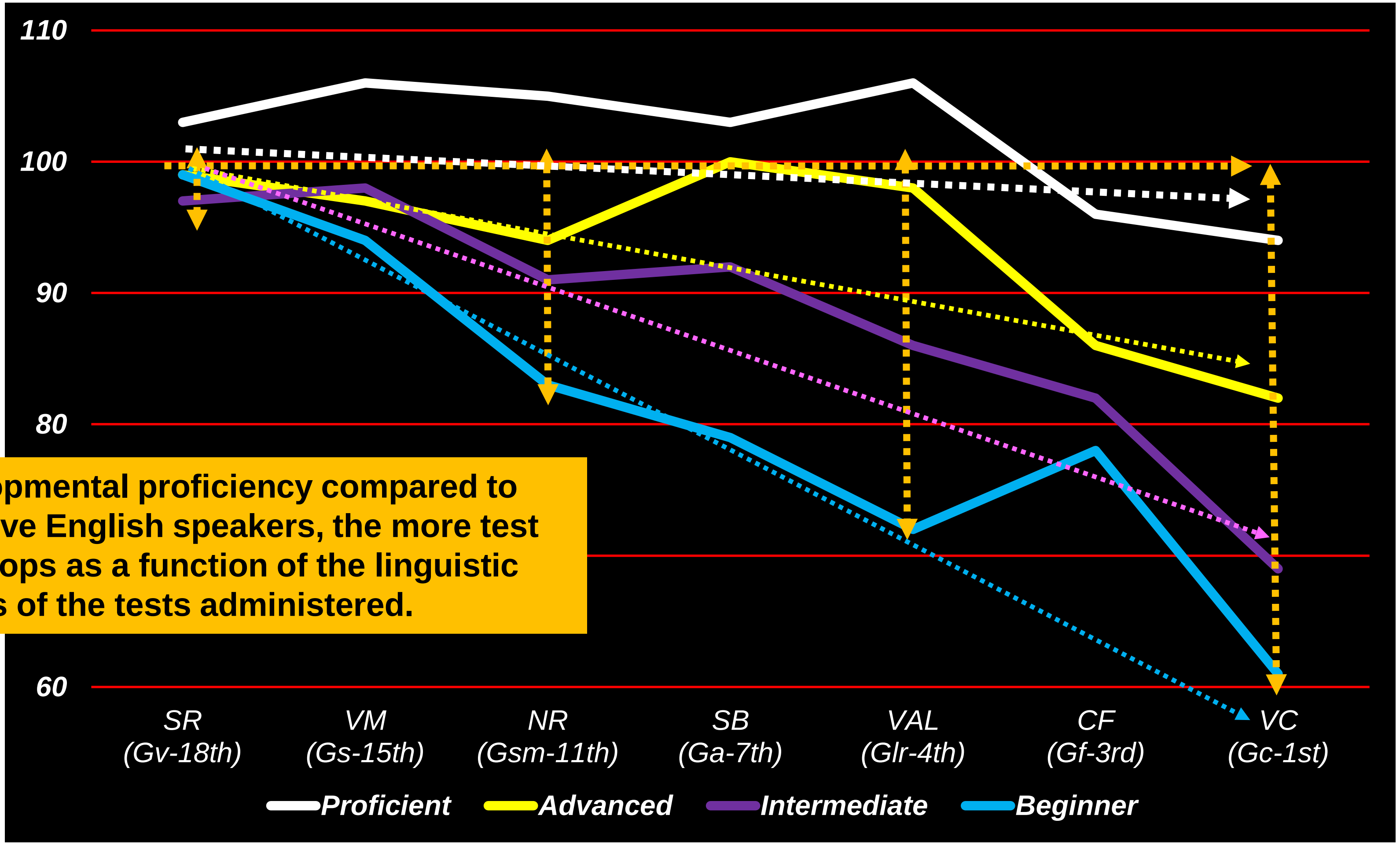
"[Lifetime English Exposure] was also found to exert more influence on the variance of the raw scores on the Ortiz PVAT compared to age…and because the Ortiz PVAT measures receptive language, or specifically receptive vocabulary, in English, the strong effect of Lifetime English Exposure above and beyond age was observed (pp. 51).

Source: Wong, J. Y. T. (2023). On the Importance of True Peer Norms in the Assessment of English Learners: A Validation Study of the Ortiz Picture Vocabulary Acquisition Test. Doctoral dissertation, St. John's University, Jamaica, Queens, NY.

Research Foundations for ML Evaluation

Research Principle 3: ML performance is moderated by linguistic/acculturative variables

Domain specific scores across the seven WJ III subtests according to language proficiency level on the NYSESLAT



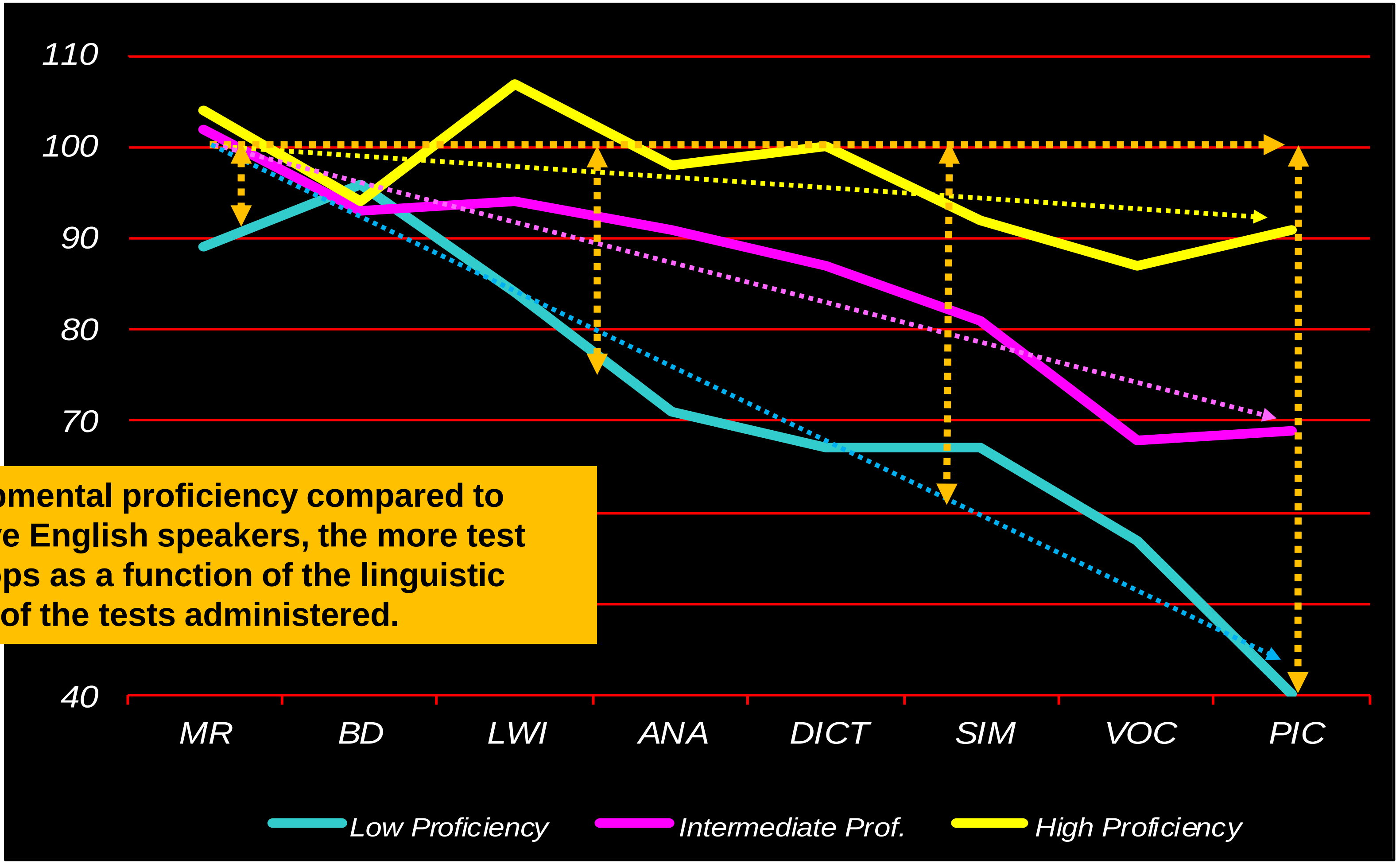
The less developmental proficiency compared to monolingual native English speakers, the more test performance drops as a function of the linguistic demands of the tests administered.

Source: Sotelo-Dynega, M., Ortiz, S.O., Flanagan, D.P., Chaplin, W. (2013). English Language Proficiency and Test Performance: Evaluation of bilinguals with the Woodcock-Johnson III Tests of Cognitive Ability. Psychology in the Schools, Vol 50(8), pp. 781-797.

Research Foundations for ML Evaluation

Research Principle 3: ML performance is moderated by linguistic/acculturative variables

Mean subtest scores across the four WASI subtests and four WMLS-R subtests according to language proficiency level



The less developmental proficiency compared to monolingual native English speakers, the more test performance drops as a function of the linguistic demands of the tests administered.

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

A Critical Review of C-LIM Research: Styck & Watkins

The main finding in the study is stated as follows:

> *"The valid C-LIM profile (i.e., cell means did not decline) emerged in the mean WISC-IV normative sample and the ELL sample." (p. 374). (emphasis added)*

It is clear that the normative sample "did not decline" as their mean on every subtest was invariant,10.3 (SS=102). However, for the EL sample, the highest mean was on Picture Concepts (SS=98) and lowest was on Vocabulary (SS=85). With minor variation, examination of the data in the following table strongly suggests a clear decline in the EL sample's means.

*Source: Styck, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382.

A Critical Review of C-LIM Research: Styck & Watkins

Decline or No Decline? Comparison of Means for WISC-IV Subtests

WISC-IV Subtest	Norm Sample Mean[a]	ELL Mean 2013	Difference[b]	ELL Mean 2014	Difference[b]
Picture Concepts	102	98	4	94	8
Matrix Reasoning	102	96	6	93	9
Symbol Search	102	95	7	93	9
Block Design	102	94	8	93	9
Coding	102	94	8	92	10
Comprehension	102	92	10	88	14
Letter-Number Sequencing	102	88	14	84	18
Similarities	102	88	14	86	16
Digit Span	102	87	15	84	14
Vocabulary	102	85	17	82	20

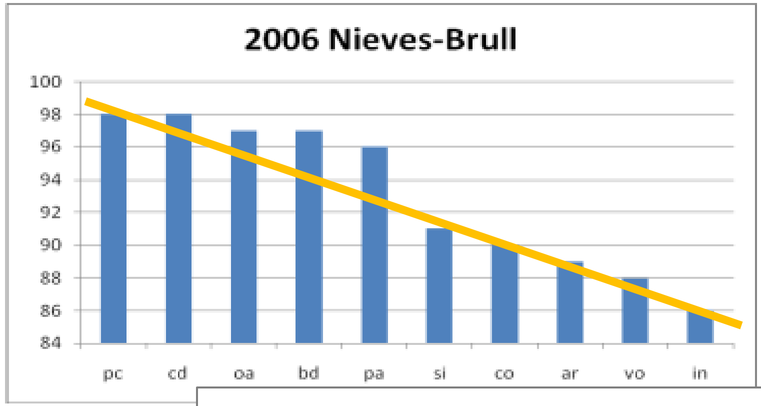
[a] *Means were reported in the study as Scaled Scores (e.g., 10.3). They have been converted here to Deviation IQ metric for the sake of simplicity.*

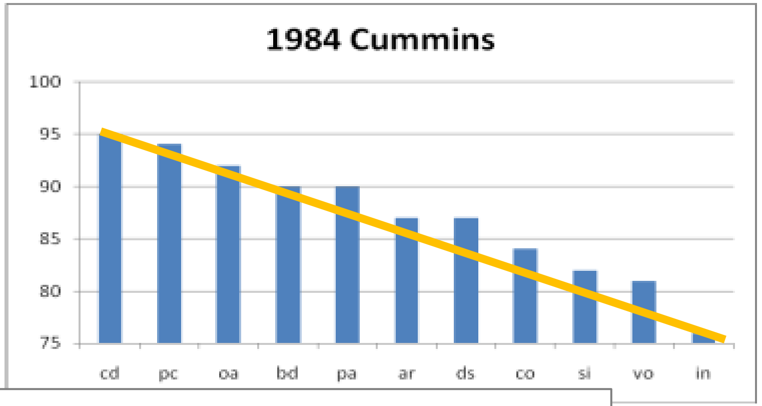
[b] *The difference between all 15 norm sample and ELL subtest and composite means were found to be statistically significant at the p<.001 level.*

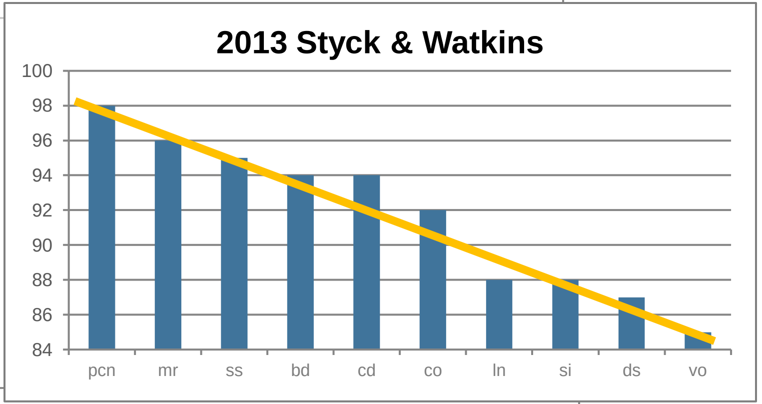
*Source: Styck, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382.

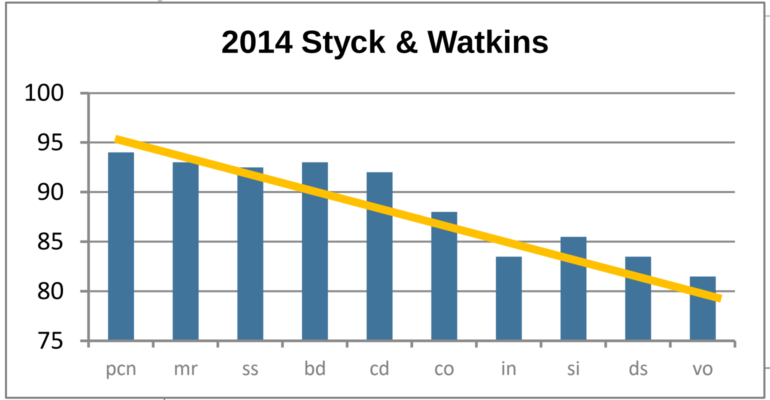
A Critical Review of C-LIM Research: Styck & Watkins

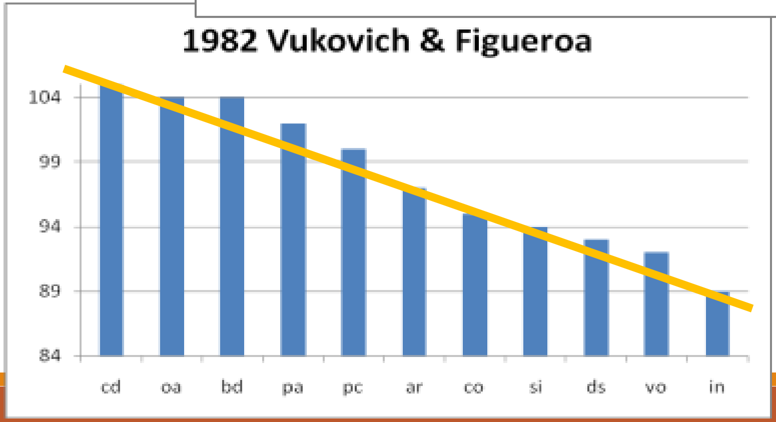
Comparison of 2013/2014 Styck & Watkins data and other WISC studies with ELs

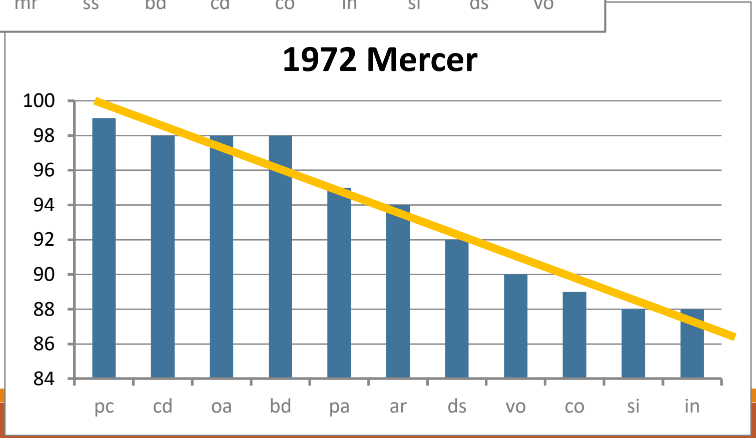












A Critical Review of C-LIM Research: Styck & Watkins

Main conclusion in the study is stated as follows:

"Thus, neither sample of children exhibited the invalid C-LIM profile when group mean scores were considered" (p. 374) (emphasis added).

The "invalid C-LIM profile" would be indicated by a systematic decline in mean scores in the matrix meaning that the test results were influenced primarily by the presence of cultural and linguistic variables.

The C-LIM is intended to compare individual performance against the group, not evaluate group scores, especially from a population where 97% have identified disabilities. Nevertheless, with a sufficiently large sample such differences in performance are likely to become more and more randomly distributed. Moreover, the C-LIM is certainly subject to modification on the basis of additional quality research.

*Source: Styck, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382.

A Critical Review of C-LIM Research: Styck & Watkins

Comparison of Order of Means for WISC-IV Classifications for ELL Group

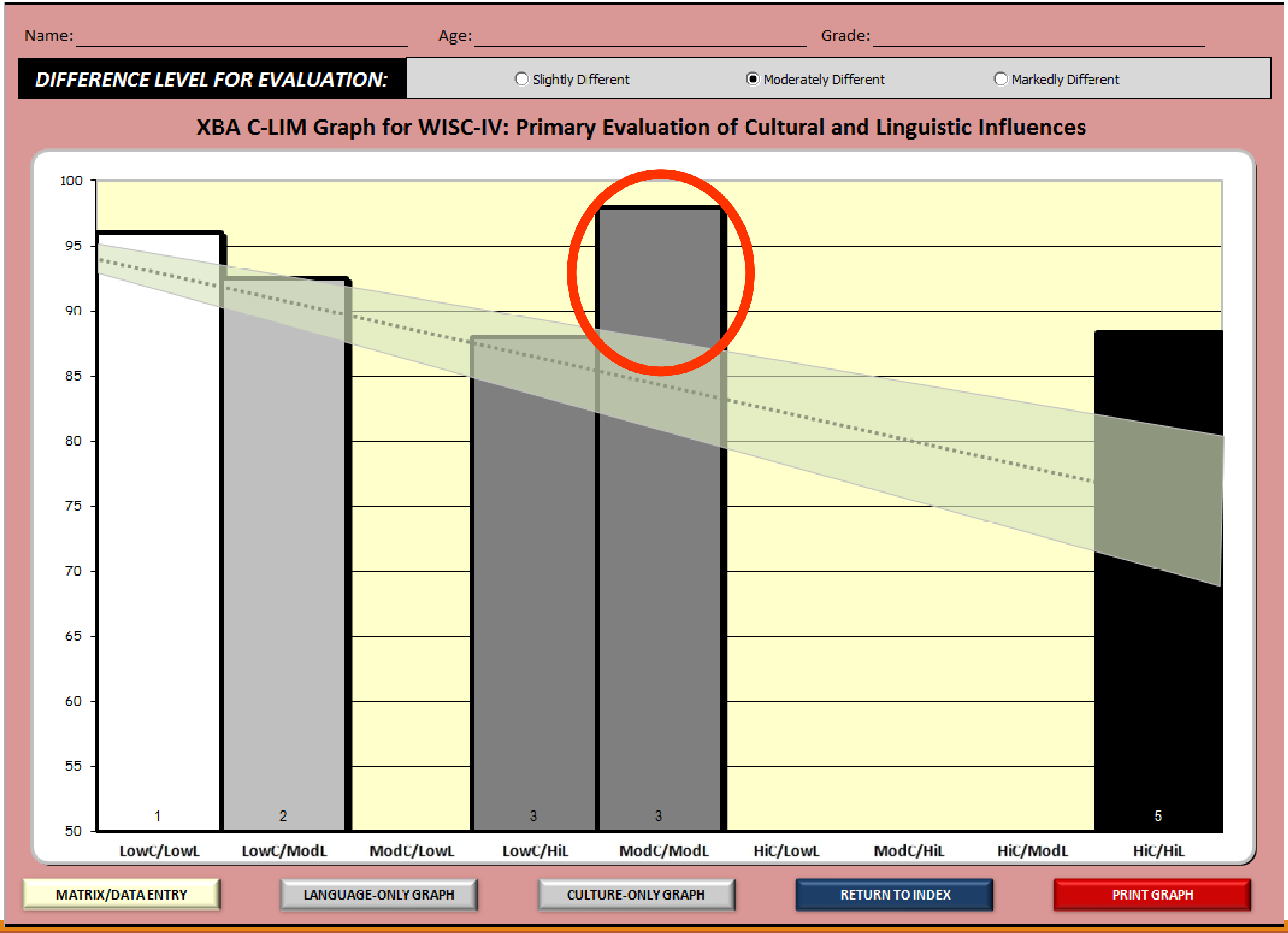
	C-LIM Classifications	Styck and Watkins, 2013*	Subtest Means
Tier 1	Matrix Reasoning	Picture Concepts Matrix Reasoning	98 96
Tier 2	Symbol Search Block Design Coding Digit Span	Symbol Search Block Design Coding Comprehension	95 94 94 92
Tier 3	Letter-Number Sequencing Picture Concepts	Letter-Number Sequencing	88
Tier 4			
Tier 5	Similarities Comprehension Vocabulary	Similarities Digit Span Vocabulary	88 87 85

7 of the 10 WISC-IV subtest means follow the exact C-LIM classifications

*Source: Styck, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382.

A Critical Review of C-LIM Research: Styck & Watkins

WISC-IV DATA FOR ELL GROUP WITH ORIGINAL CLASSIFICATIONS (ENGLISH)



A Critical Review of C-LIM Research: Styck & Watkins

Styck and Watkins interpreted the C-LIM backwards and this is what passes as peer review!

The study noted that:

"roughly 97% of (n = 83) of participants were identified as meeting criteria for an educational disability (86% as SLD)" (p. 371).

As noted previously, this suggests that individual C-LIM profiles should display ***valid*** results, not invalid, since valid results are needed to support the district's identification of a disability.

When individual C-LIM's for the ELL group were examined, they found that nearly 89.5% of the ELLs did in fact display valid results indicating that any low scores could well reflect a disability and indicating a very high degree of consistency with the clinical decisions made by the district's eligibility team.

*Source: Styck, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382.

A Critical Review of C-LIM Research: Styck & Watkins

Overall decline ***and*** within expected range = no disability

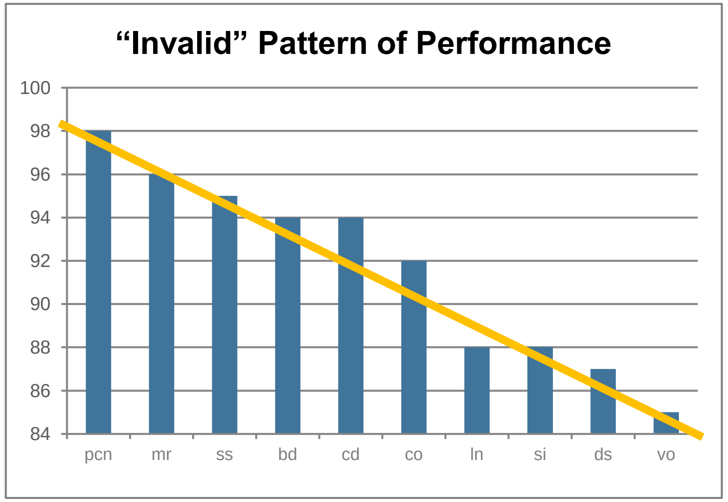
No decline ***or*** below expected range = possible disability

WISC-IV C-LIM Analysis		EL Sample (with disability)	Norm Sample (no disability)
	Invalid Scores (decline)	N=9 (N=6, 7.0%) (N=3, 3.5%)	N = 100 (4.9%)
	Valid Scores (no decline)	N = 77 (89.5%)	N = 1,933 (95.1%)

The most egregious error in the Styck & Watkins studies is found in the examination of *individual* patterns of performance within the C-LIM where they expected "invalid" (declining performance) instead of "valid" (non-declining performance) given that their nearly their entire sample had already been identified as having a disability. The authors noted that *"roughly 97% of (n = 83) of participants were identified as meeting criteria for an educational disability (86% as SLD)" (p. 371).* Yet only 9 ELL cases (10.5%) resulted in invalid scores (no disability). Thus, the C-LIM suggested invalid scores in 9 cases, 3 of which were likely correct (those without disabilities) indicating that the C-LIM was consistent with and supported the placement decision of the child by the district in 93% of the cases (89.5% + 3.5%). Moreover, the results of analyses with the WISC-IV normative sample show that declines relative to language are unusual, perhaps even indications of potential SLI in monolingual, native English speakers as described by Cormier et al. (2014).

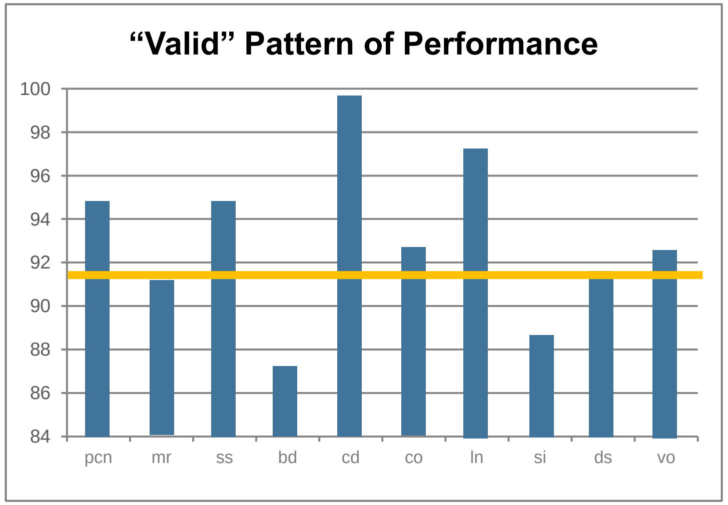
*Table adapted from: Styck, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382.

A Critical Review of C-LIM Research: Styck & Watkins



Invalid Pattern	Expected N = 3 (out of 86)
No evidence of disability	Correct (N = 3, 3.5%) Incorrect (N = 6, 7.0%)

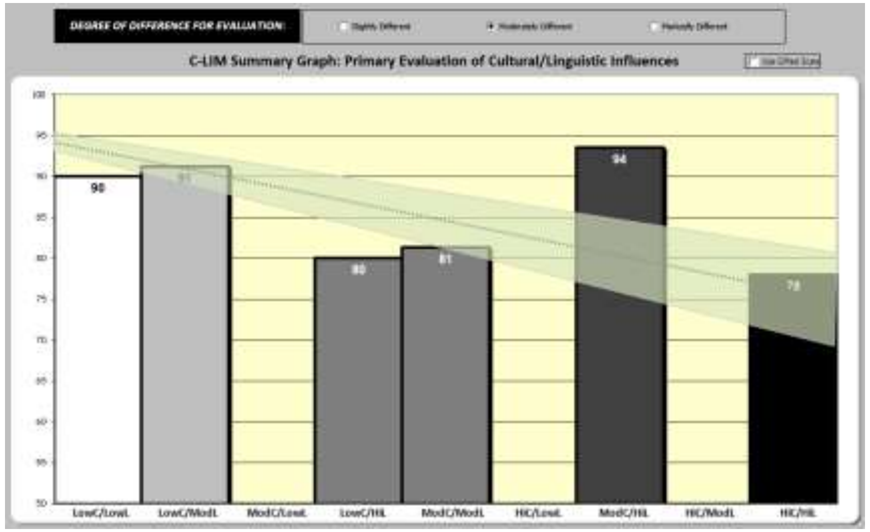
Correct C-LIM pattern found in 89.5% + 3.5% = **93%** of all cases

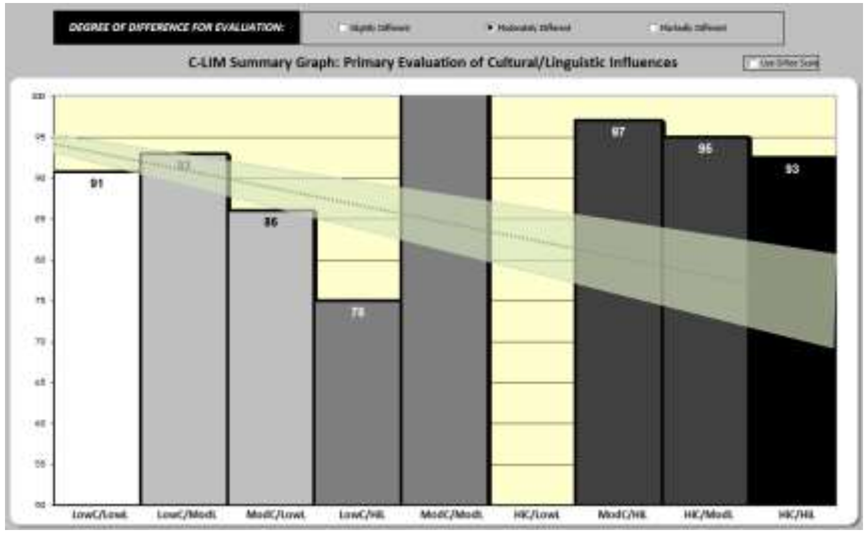


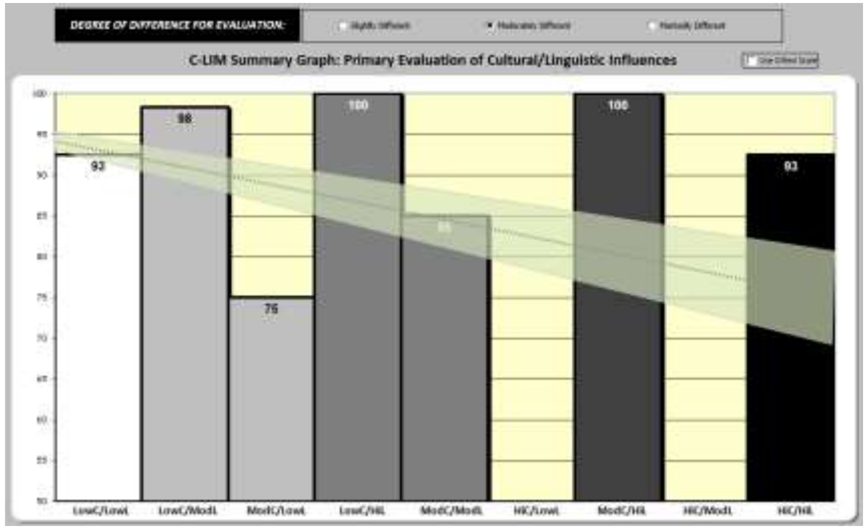
Valid Pattern	Expected N = 83 (out of 86)
Evidence of disability	Correct (N = 77, 89.5%)

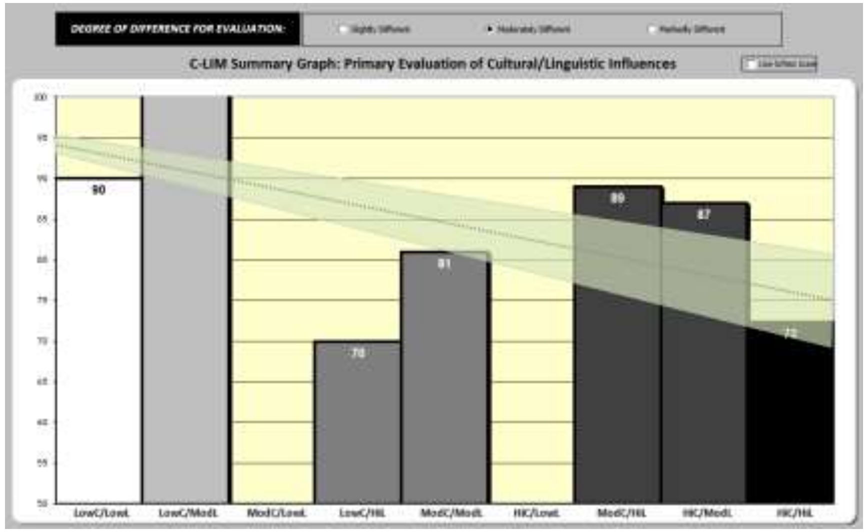
Far from undermining the validity of the C-LIM, the Styck & Watkins studies provide powerful support for the clinical utility and validity of the C-LIM when evaluating EL test performance using current research and an evidence-base.

A Critical Review of C-LIM Research: Styck & Watkins: Which one is the ML?









When 97% of your sample already possesses a disability, patterns for English speakers and Multilingual Learners are indistinguishable. The C-LIM is NOT an ML identification test.

A Critical Review of C-LIM Research: Kranzler et al.

Badly designed research conducted with limited understanding of the variables involved is not evidence of a problem in current knowledge, it's more a reflection of poor research.

According to the demographic information regarding the participants used in the study, the sample:

1. was comprised of ELLs with a mean age of 11 with an average grade placement in 6th (i.e., already learned to read/write and do math)

2. of the included ELLs, approximately 74% had been educated in their native language and country prior to coming to the U.S. (i.e., had CALP)

3. was extremely small (n=46) and no measures of proficiency (i.e., no control for developmental differences in the heritage or English language)

Thus, the age, grade, and background of 3/4th of the ELs in the sample indicated that they had already acquired mature and fluent academic skills in their heritage language (i.e., had developed CALP) prior to starting education in the U.S. Cummins' linguistic transfer model would predict better cognitive performance as compared to ELs who began school in the U.S. without heritage language instruction. But, despite these major differences, the results remained consistent with the research underlying the C-LIM, especially that indicated by the "slightly different range.

A Critical Review of C-LIM Research: Kranzler et al.

WJ III DATA FOR PARTICIPANTS IN STUDY (ENGLISH)

XBA Culture-Language Interpretive Matrix (XBA C-LIM v2.0) for WJ III NU COG

Name: *Kranzler et al. Sample* Age: ______ Grade: ______

CLEAR DATA | SAVE DATA

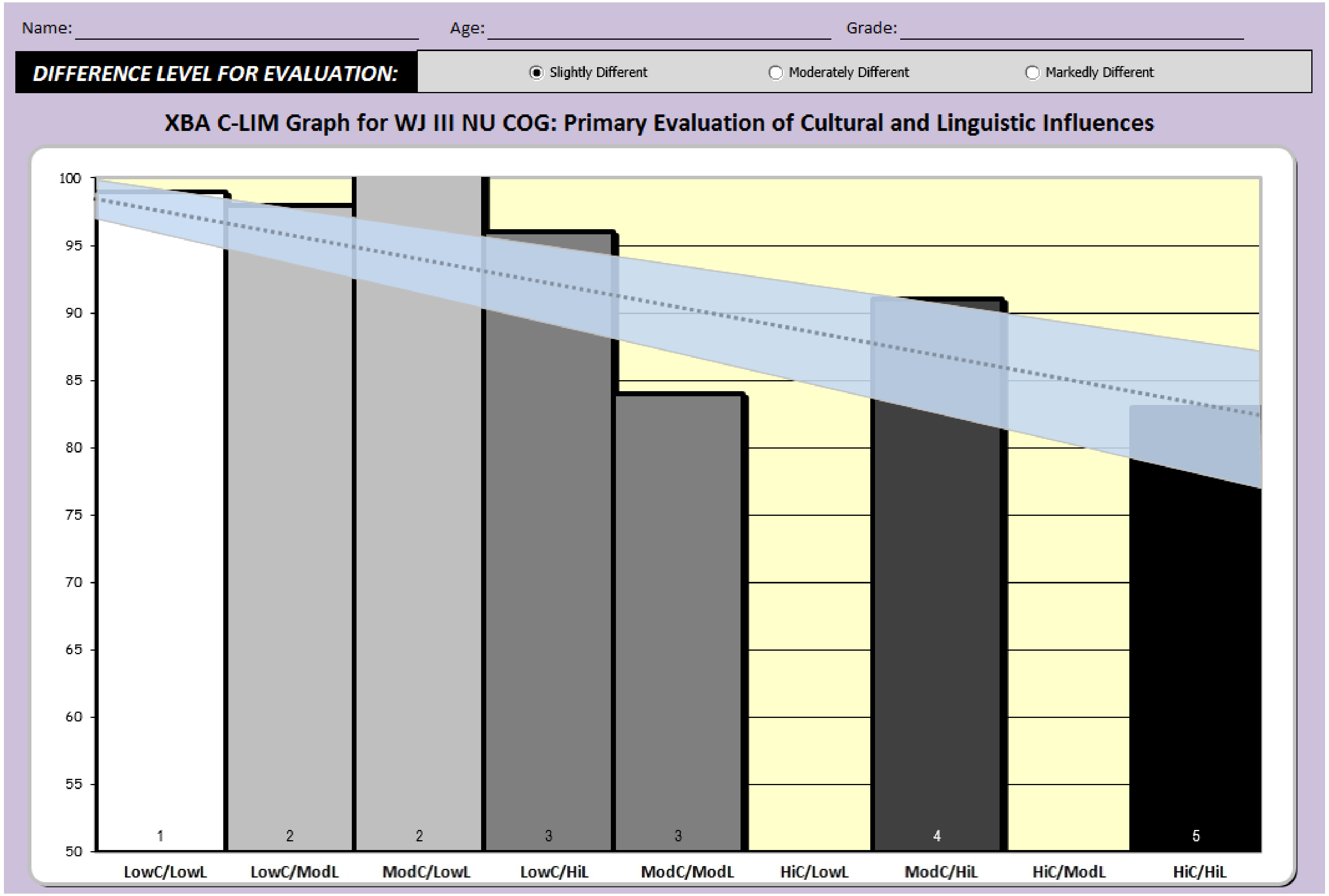
DEGREE OF LINGUISTIC DEMAND

DEGREE OF CULTURAL LOADING	LOW	Score		MODERATE	Score		HIGH	Score	
LOW	WJ III Spatial Relations	99	*99*	WJ III Numbers Reversed	99	*99*	WJ III Analysis-Synthesis		
				WJ III Visual Matching	97	*97*	WJ III Auditory Working Memory		
							WJ III Concept Formation	96	*96*
	Cell Average =	99		Cell Average =	98		Cell Average =	96	
MODERATE	WJ III Pair Cancellation			WJ III Delayed Recall: Visual Auditory Learning			WJ III Auditory Attention		
	WJ III Picture Recognition	102	*102*	WJ III Rapid Picture Naming			WJ III Decision Speed		
	WJ III Planning			WJ III Retrieval Fluency			WJ III Incomplete Words		
				WJ III Visual Auditory Learning	84	*84*	WJ III Memory for Words		
							WJ III Sound Blending	91	*91*
	Cell Average =	102		Cell Average =	84		Cell Average =	91	
HIGH							WJ III General Information		
							WJ III Verbal Comprehension	83	*83*
	Cell Average =			Cell Average =			Cell Average =	83	

CULTURE-LANGUAGE GRAPH | LANGUAGE-ONLY GRAPH | CULTURE-ONLY GRAPH | RETURN TO INDEX | PRINT MATRIX

A Critical Review of C-LIM Research: Kranzler et al.

WJ III DATA FOR PARTICIPANTS IN STUDY (ENGLISH)



A Critical Review of C-LIM Research: Kranzler et al.

Comparison of Order of Means for WJ III Classifications

	C-LTC Classifications	Kranzler et al., 2010*
Level 1	Gv - Spatial Relations	Gv - Spatial Relations
Level 2	Gsm - Numbers Reversed	Gsm - Numbers Reversed
	Gs - Visual Matching	Gs - Visual Matching
Level 3	Gf - Concept Formation	Gf - Concept Formation
Level 4	Glr - Visual Auditory Learning	Ga - Sound Blending
	Ga - Sound Blending	Glr - Visual Auditory Learning
Level 5	Gc - Verbal Comprehension	Gc - Verbal Comprehension

*Source: Kranzler, J., Flores, C., & 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, 2010, 39(3), 431-446.

A Critical Review of C-LIM Research: Kranzler et al.

Badly designed research conducted with limited understanding of the variables involved is not evidence of a problem in current knowledge, it's more a reflection of poor research.

Inference for Means: Comparing Two Independent Samples

(To use this page, your browser must recognize JavaScript.)

Choose which calculation you desire, enter the relevant population values for mu1 (mean of popu
power, a sample size (assumed the same for each sample). You may also modify α (type I error ra

- ○ Calculate Sample Size (for specified Power)
- ◉ Calculate Power (for specified Sample Size)

Enter a value for mu1: 100
Enter a value for mu2: 97
Enter a value for sigma: 15

- ◉ 1 Sided Test
- ○ 2 Sided Test

Enter a value for α (default is .05): .05
Enter a value for desired power (default is .80): 0.25
The sample size (for each sample separately) is: 46

Calculate

Reference: The calculations are the customary ones based on normal distributions. See for examp
Comparing Two Means in Bernard Rosner's **Fundamentals of Biostatistics**.

Rollin Brant
Email me at: *rollin@stat.ubc.ca*

Inference for Means: Comparing Two Independent Samples

(To use this page, your browser must recognize JavaScript.)

Choose which calculation you desire, enter the relevant population values for mu1 (mean of popu
power, a sample size (assumed the same for each sample). You may also modify α (type I error ra

- ◉ Calculate Sample Size (for specified Power)
- ○ Calculate Power (for specified Sample Size)

Enter a value for mu1: 100
Enter a value for mu2: 97
Enter a value for sigma: 15

- ◉ 1 Sided Test
- ○ 2 Sided Test

Enter a value for α (default is .05): .05
Enter a value for desired power (default is .80): 0.80
The sample size (for each sample separately) is: 310

Calculate

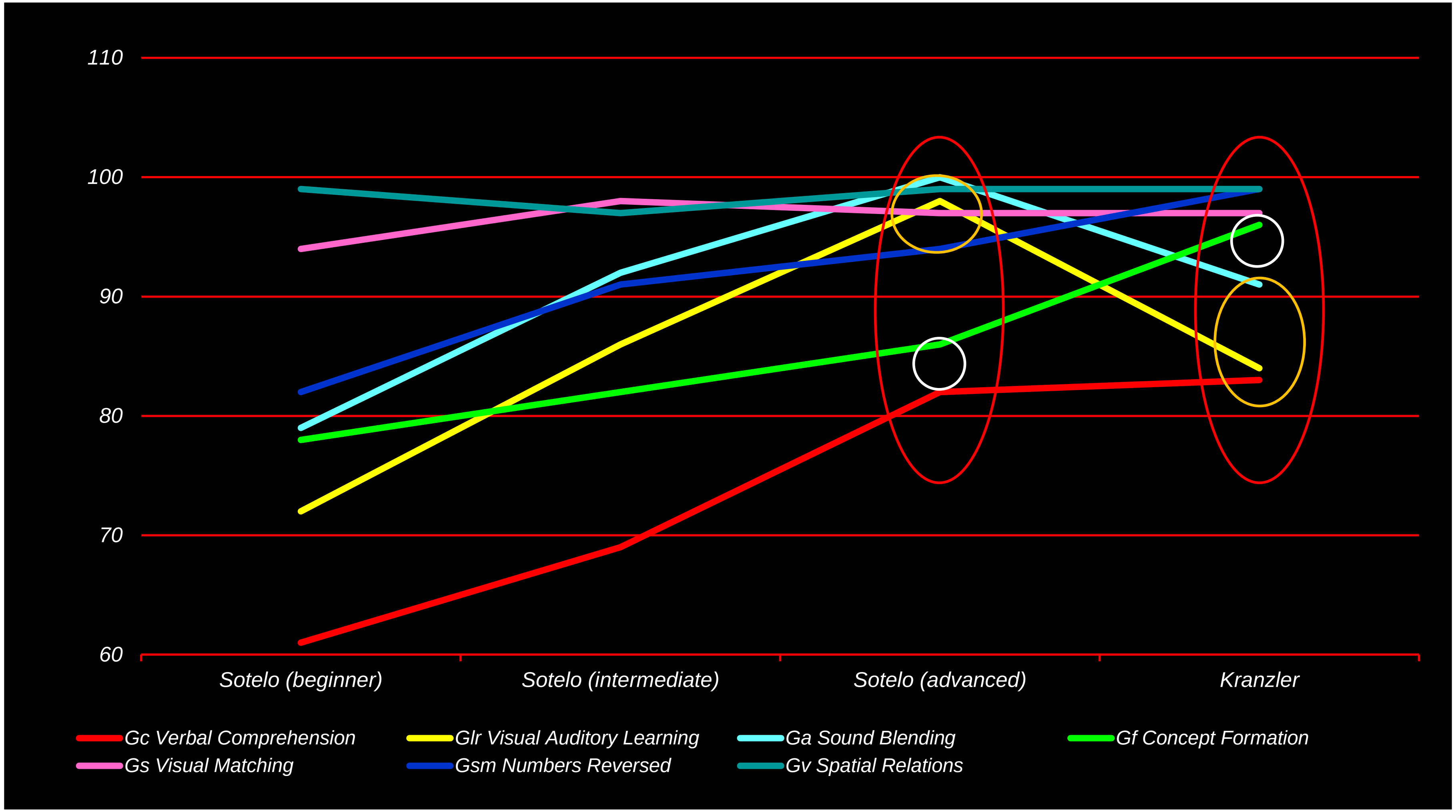
Reference: The calculations are the customary ones based on normal distributions. See for examp
Comparing Two Means in Bernard Rosner's **Fundamentals of Biostatistics**.

Rollin Brant
Email me at: *rollin@stat.ubc.ca*

To detect an 8-point difference with default power (.80), requires a sample size of: n=44
To detect a 5-point difference with default power (.80), requires a sample size of: n=112
To detect a 4-point difference with default power (.80), requires a sample size of: n=174
To detect a 3-point difference with default power (.80), requires a sample size of: n=310

A Critical Review of C-LIM Research: Kranzler et al.

Mean subtest scores across the seven WJ III subtests – Comparison of Sotelo-Dynega and Kranzler et al. Data



Source: Kranzler, J., Flores, C., & 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, 2010, 39(3), 431-446.

A Critical Review of C-LIM Research: Kranzler et al.

Results of the Kranzler et al. study indicated that:

1. Despite use of an EL sample that was older, and which had been educated before coming to the U.S., the overall results still showed a decline in performance as tests become more culturally/linguistically bound, just less so.
2. All WJIII subtest mean values for the ELL sample, which was a non-referred, non-disabled sample, were within the C-LIM "slightly different" range or higher.
3. Despite a very small sample size and limited test administration (8 subtests only), the order of decline for subtest means from the WJIII are nearly identical to the order as indicated by the classification of WJ III tests within the C-LIM

The bottom line: Kranzler et al. concluded that: *"a statistically significant (decreasing) trend was observed for the effect of linguistic demand and cultural loading combined."*

Their criticism of the C-LIM was based on inappropriate expectations of precision (differences too small to be detectable by their sample size) and a pattern of decline that simply was not consistent with their population (a higher functioning, "slightly different," group sample). Nevertheless, the results provide considerable support for the WJIII classifications within the C-LIM *and what should be criticized is the lack of quality of the research in failing to account for developmental language issues for ELs at various ages and grades*.

Research Foundations of the C-LIM and its Validity

Basic C-LIM Research Principle: Language proficiency moderates test score performance proportionally.

"One such challenge is assessing the cognitive abilities of the growing number of students who are considered ELs; limited English proficiency can lead to linguistically biased test results, which would lead to a misrepresentation of the examinee's true cognitive abilities.

To eliminate this potential source of bias, psychologists testing EL students could consider examinee characteristics before administering a standardized measure of cognitive ability.

This idea is not new. More than a decade ago, Flanagan et al. (2007) noted the critical need for psychologists to collect information regarding students' level of English proficiency, and the level of English required for the student to be able to comprehend test directions, formulate and communicate responses, or otherwise use their English language abilities within the testing process.

Nonetheless, the results of our study provide an empirical basis in support of this broad recommendation." (p. 9)

Source: Cormier, D. C., Bulut, O., McGrew, K. S. & Kennedy, K. (2022). Linguistic Influences on Cognitive Test Performance: Examinee Characteristics Are More Important than Test Characteristics, Journal of Intelligence, Volume 10, Issue 1.

Conclusions and Comments

Why are we so quick to believe that a few misguided studies are enough to overturn over a century's worth of data that show that MLs perform lower on tests compared to monolinguals?

- Science is a battle for what the "truth" is and the issues that affect this battle are no different than those found in general society, including deliberate oppression and systemic racism.
- Sensationalism borrowed from social media is now standard in academic journals and scientific discussions, and that which is titillating and controversial drives the most interest.
- Making one's scholarly reputation in the field never comes from agreeing with what is considered "truth" but rather by challenging that truth even if the challenge is unfounded.
- The absence of evidence is not the same as evidence of absence.
- Garbage in, garbage out.

For someone to discredit the validity of the C-LIM, all they need do is show that evaluation of a wide range of abilities in monolinguals and bilinguals with varying levels of English proficiency does NOT show any differences in performance, even on verbal tests. This will never happen.