Evidence-based Evaluation of English Learners:

A contemporary approach to testing.

Oklahoma School Psychologists Association

Tulsa, OK

March 30, 2018

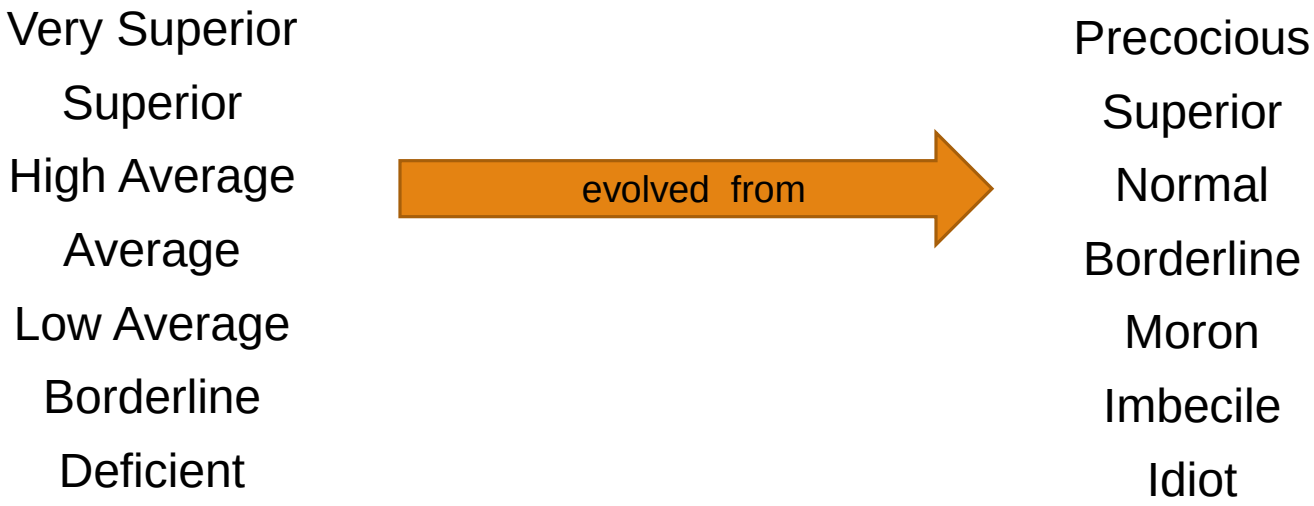
Samuel O. Ortiz, Ph.D.
St. John's University

The Testing of Bilinguals: Early influences and a lasting legacy.

It was believed that:

- *speaking English, familiarity with and knowledge of U.S. culture had no bearing on intelligence test performance*
- *intelligence was genetic, innate, static, immutable, and largely unalterable by experience, opportunity, or environment*
- *being bilingual resulted in a "mental handicap" that was measured by poor performance on intelligence tests and thus substantiated its detrimental influence*

Much of the language and legacy ideas remain embedded in present day tests.



The Testing of Bilinguals: Early influences and a lasting legacy.

H. H. Goddard and the menace of the feeble-minded

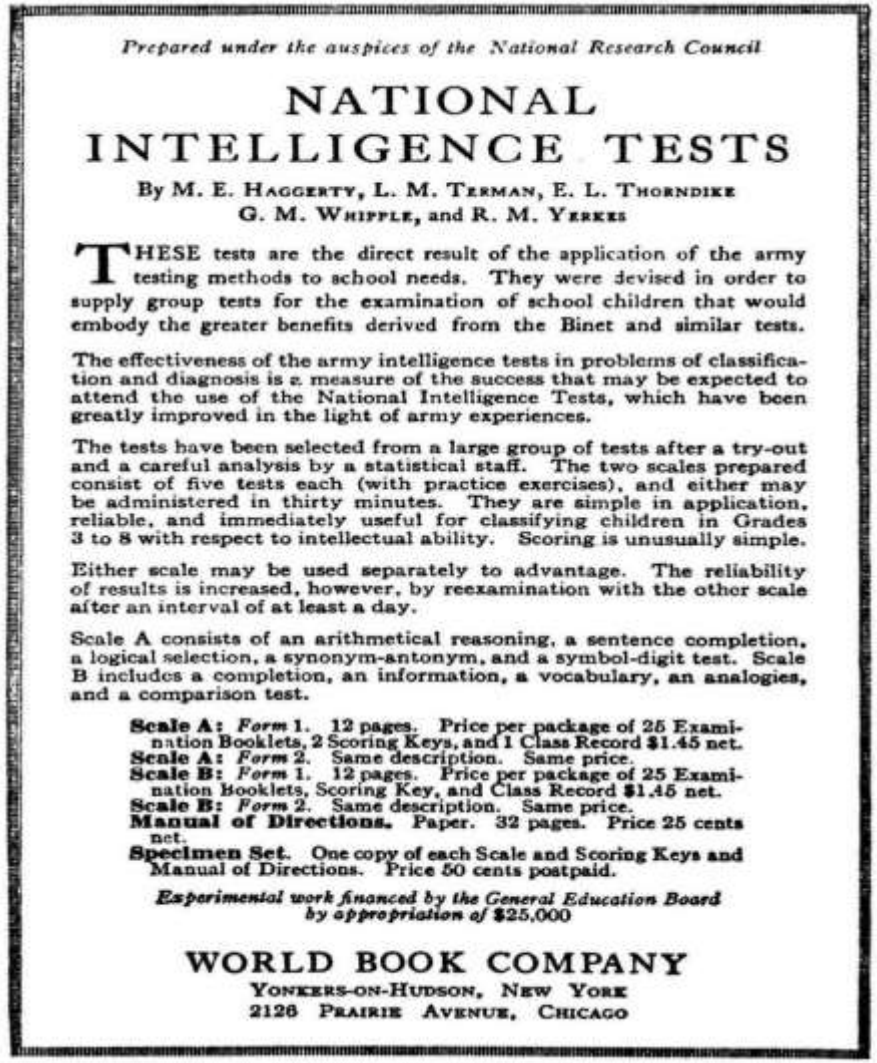
- *The testing of newly arrived immigrants at Ellis Island*

Lewis Terman and the Stanford-Binet

- *America gives birth to the IQ test of inherited intelligence*

Robert Yerkes and mass mental testing

- *Emergence of the bilingual-ethnic minority "handicap"*



Prepared under the auspices of the National Research Council

NATIONAL INTELLIGENCE TESTS

By M. E. Haggerty, L. M. Terman, E. L. Thorndike
G. M. Whipple, and R. M. Yerkes

THESE tests are the direct result of the application of the army testing methods to school needs. They were devised in order to supply group tests for the examination of school children that would embody the greater benefits derived from the Binet and similar tests.

The effectiveness of the army intelligence tests in problems of classification and diagnosis is a measure of the success that may be expected to attend the use of the National Intelligence Tests, which have been greatly improved in the light of army experiences.

The tests have been selected from a large group of tests after a try-out and a careful analysis by a statistical staff. The two scales prepared consist of five tests each (with practice exercises), and either may be administered in thirty minutes. They are simple in application, reliable, and immediately useful for classifying children in Grades 3 to 8 with respect to intellectual ability. Scoring is unusually simple.

Either scale may be used separately to advantage. The reliability of results is increased, however, by reexamination with the other scale after an interval of at least a day.

Scale A consists of an arithmetical reasoning, a sentence completion, a logical selection, a synonym-antonym, and a symbol-digit test. Scale B includes a completion, an information, a vocabulary, an analogies, and a comparison test.

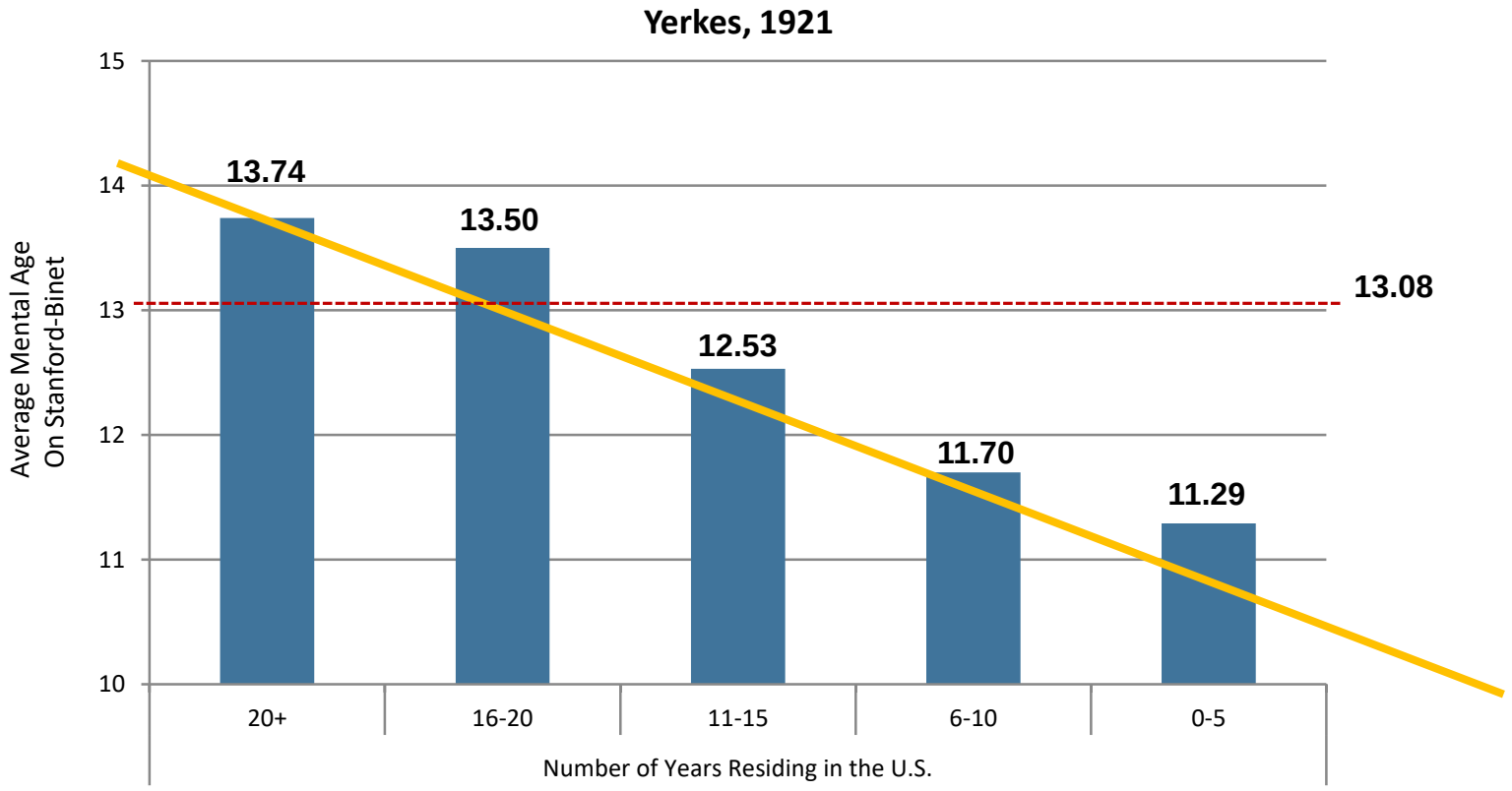
Scale A: *Form* 1. 12 pages. Price per package of 25 Examination Booklets, 2 Scoring Keys, and 1 Class Record $1.45 net.
Scale A: *Form* 2. Same description. Same price.
Scale B: *Form* 1. 12 pages. Price per package of 25 Examination Booklets, Scoring Key, and Class Record $1.45 net.
Scale B: *Form* 2. Same description. Same price.
Manual of Directions. Paper. 32 pages. Price 25 cents net.
Specimen Set. One copy of each Scale and Scoring Keys and Manual of Directions. Price 50 cents postpaid.

Experimental work financed by the General Education Board by appropriation of $25,000

WORLD BOOK COMPANY
Yonkers-on-Hudson, New York
2126 Prairie Avenue, Chicago

Early Research Foundations for ELL Evaluation

Mean Mental Age (MA) from Binet Scales in a non-native English speaking sample from Yerkes' data as analyzed by C.C. Brigham (1921)



Average score for native English speakers on Beta = 101.6 (Very Superior; Grade A)

Average score for non-native English speakers on Beta = 77.8 (Average; Grade C)

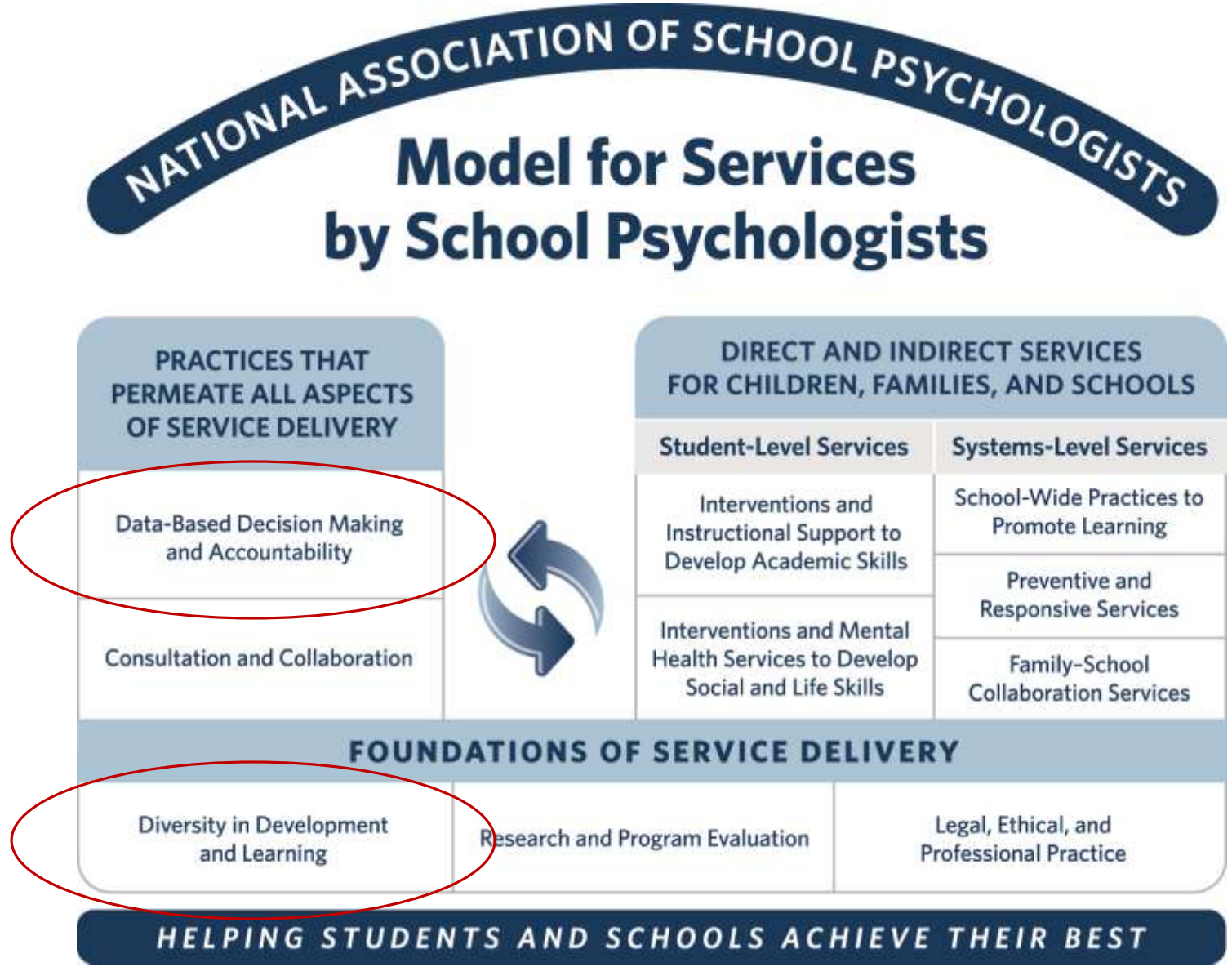
Bilingualism and Testing

- *Interpretation: New immigrants are inferior*

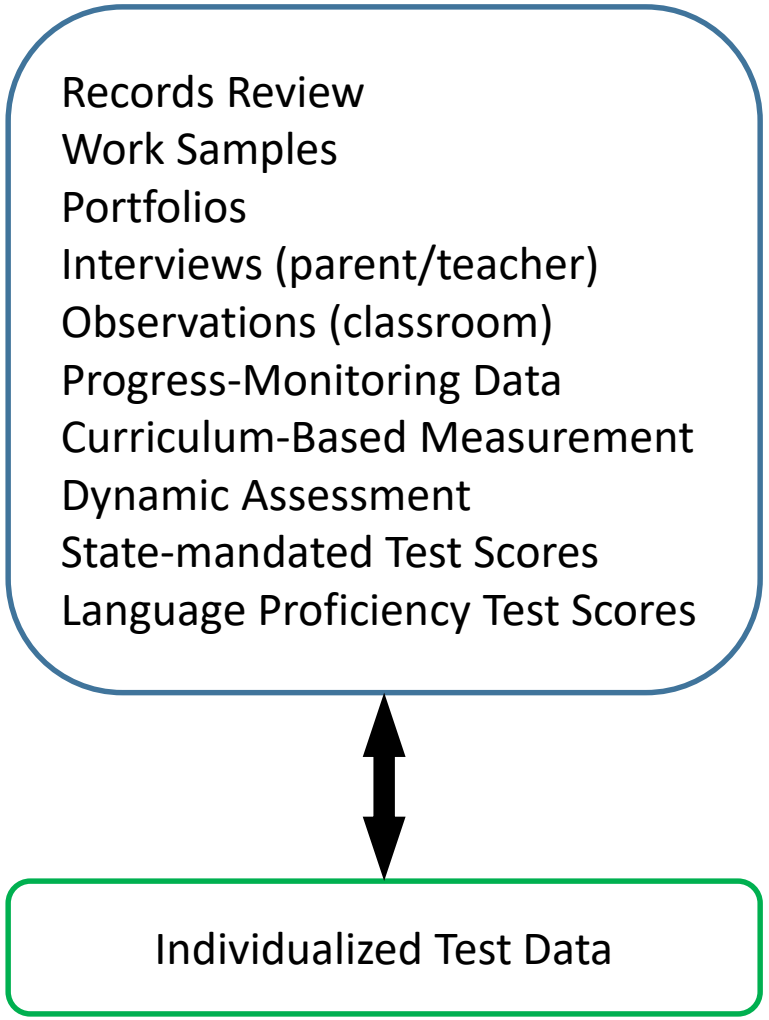
Instead of considering that our curve indicates a growth of intelligence with increasing length of residence, we are forced to take the reverse of the picture and accept the hypothesis that the curve indicates a gradual deterioration in the class of immigrants examined in the army, who came to this country in each succeeding 5 year period since 1902…The average intelligence of succeeding waves of immigration has become progressively lower.

Brigham, 1923

The NASP Practice Model



Nondiscriminatory Assessment Options



ADVANTAGES

Measures what student has actually been taught
Permits examination of learning and progress
Information is more relevant to instruction/intervention
Relies on a wide and diverse range of data/information
Avoids dealing with test validity issues

DISADVANTAGES

Standard for true peer comparison remains problematic
Not many training programs teach dynamic assessment
Eligibility criteria may be difficult to demonstrate
Difficult to ascertain "average" unless can use local norms
Requires significant foundational knowledge of issues

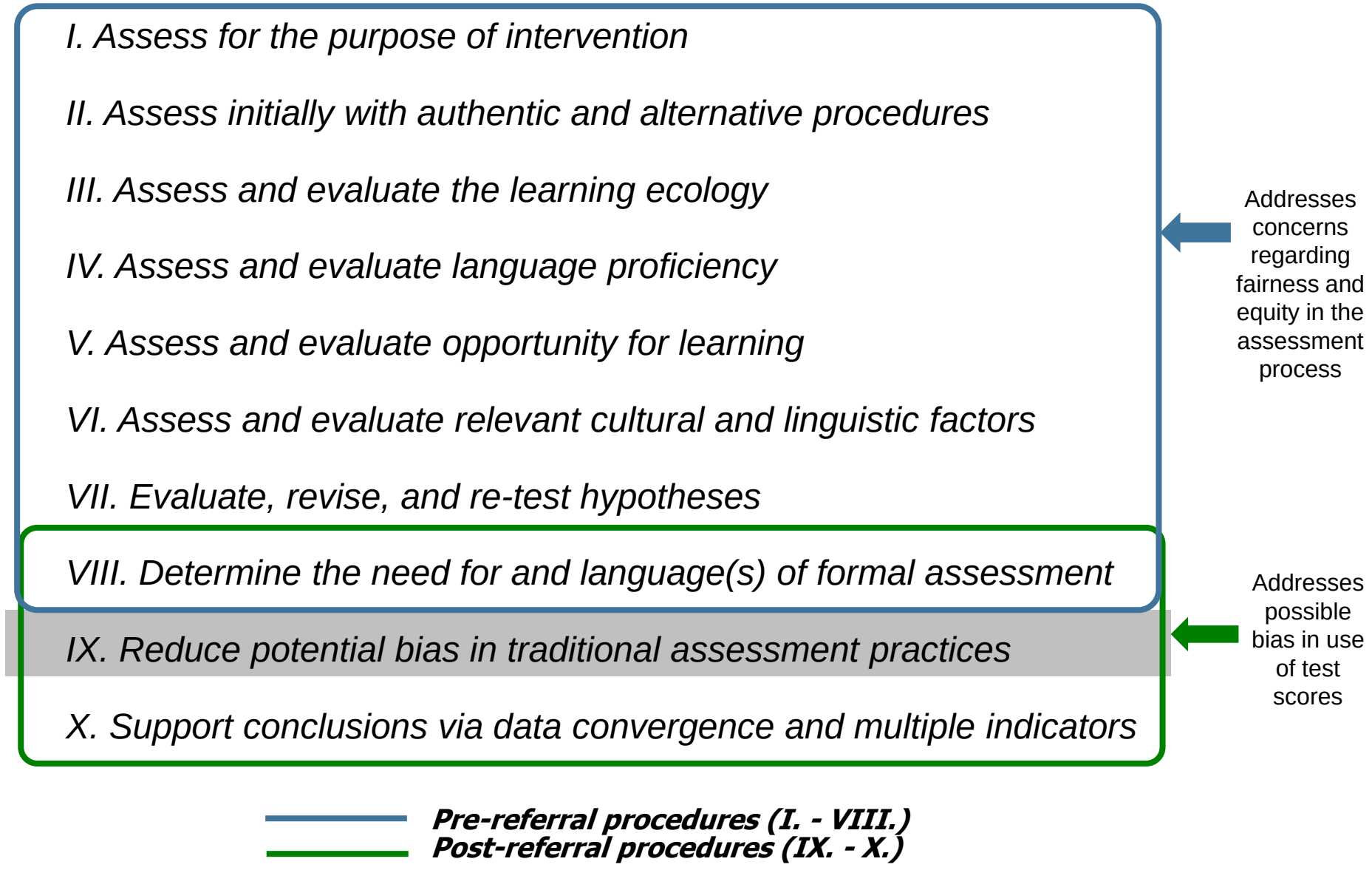
ADVANTAGES

Measures what student can actually compared to others
Eligibility criteria easier to demonstrate with numbers

DISADVANTAGES

Standard for true peer comparison remains problematic
Requires significant foundational knowledge of issues

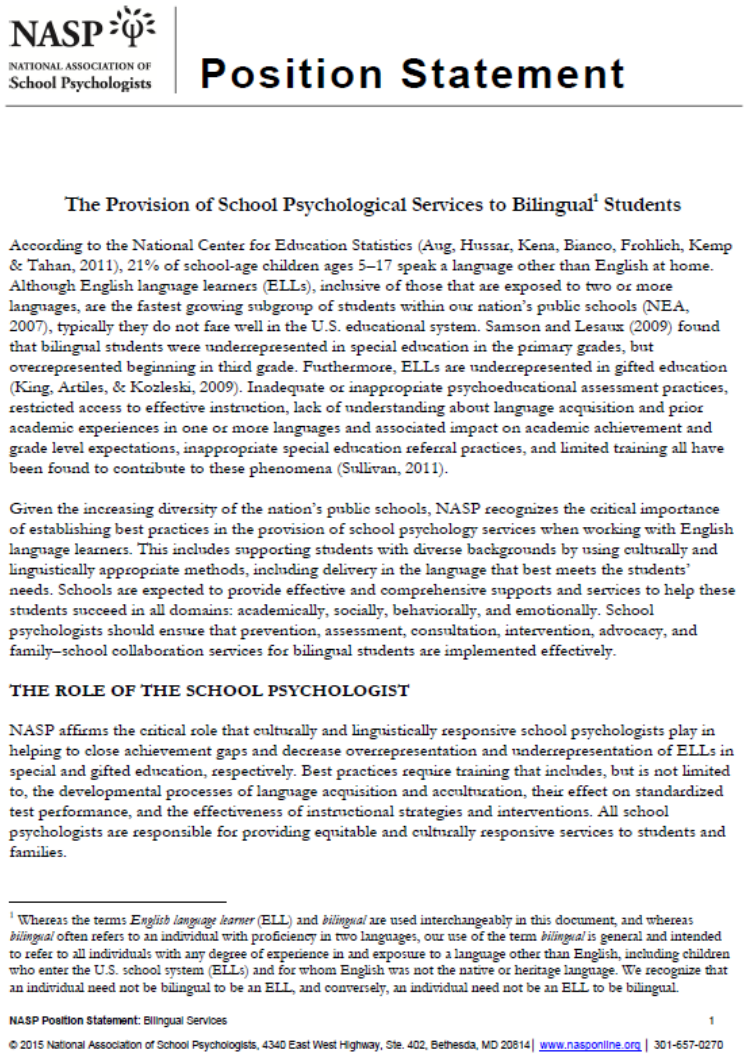
Nondiscriminatory Assessment Framework



The Provision of School Psychological Services to Bilingual Students

This document represents the very first official position by NASP on school psychology services to bilingual students was adopted in 2015.

It serves as official policy of NASP and is ***applicable to ALL school psychologists***, whether or not they are bilingual themselves.



NASP NATIONAL ASSOCIATION OF School Psychologists | **Position Statement**

The Provision of School Psychological Services to Bilingual[1] Students

According to the National Center for Education Statistics (Aud, Hussar, Kena, Bianco, Frohlich, Kemp & Tahan, 2011), 21% of school-age children ages 5–17 speak a language other than English at home. Although English language learners (ELLs), inclusive of those that are exposed to two or more languages, are the fastest growing subgroup of students within our nation's public schools (NEA, 2007), typically they do not fare well in the U.S. educational system. Samson and Lesaux (2009) found that bilingual students were underrepresented in special education in the primary grades, but overrepresented beginning in third grade. Furthermore, ELLs are underrepresented in gifted education (King, Artiles, & Kozleski, 2009). Inadequate or inappropriate psychoeducational assessment practices, restricted access to effective instruction, lack of understanding about language acquisition and prior academic experiences in one or more languages and associated impact on academic achievement and grade level expectations, inappropriate special education referral practices, and limited training all have been found to contribute to these phenomena (Sullivan, 2011).

Given the increasing diversity of the nation's public schools, NASP recognizes the critical importance of establishing best practices in the provision of school psychology services when working with English language learners. This includes supporting students with diverse backgrounds by using culturally and linguistically appropriate methods, including delivery in the language that best meets the students' needs. Schools are expected to provide effective and comprehensive supports and services to help these students succeed in all domains: academically, socially, behaviorally, and emotionally. School psychologists should ensure that prevention, assessment, consultation, intervention, advocacy, and family–school collaboration services for bilingual students are implemented effectively.

THE ROLE OF THE SCHOOL PSYCHOLOGIST

NASP affirms the critical role that culturally and linguistically responsive school psychologists play in helping to close achievement gaps and decrease overrepresentation and underrepresentation of ELLs in special and gifted education, respectively. Best practices require training that includes, but is not limited to, the developmental processes of language acquisition and acculturation, their effect on standardized test performance, and the effectiveness of instructional strategies and interventions. All school psychologists are responsible for providing equitable and culturally responsive services to students and families.

[1] Whereas the terms *English language learner* (ELL) and *bilingual* are used interchangeably in this document, and whereas *bilingual* often refers to an individual with proficiency in two languages, our use of the term *bilingual* is general and intended to refer to all individuals with any degree of experience in and exposure to a language other than English, including children who enter the U.S. school system (ELLs) and for whom English was not the native or heritage language. We recognize that an individual need not be bilingual to be an ELL, and conversely, an individual need not be an ELL to be bilingual.

Fundamental Requirements for Evaluation

According to the NASP Position Statement:

"NASP promotes the standards set by the Individuals with Disabilities Education Improvement Act (IDEA, 2004) that require the use of ***reliable and valid*** assessment tools and procedures." (p. 2; emphasis added).

NASP (2015). Position Statement: The Provision of School Psychological Services to Bilingual Students. Retrieved from http://www.nasponline.org/x32086.xml

What's the Problem with Tests and Testing with ELs?

For native English speakers, growth of cognitive abilities and knowledge acquisition are tied closely to age and assumes normal educational experiences. Thus, age-based norms effectively control for variation in development and provide an appropriate basis for comparison. However, this is not true for English learners who may neither live in a "mainstream" culture nor benefit to an equivalent degree from formal education as native English speakers.

Development Varies by Experience – Not necessarily by race or ethnicity

> *"The key consideration in distinguishing between a difference and a disorder is whether the child's performance differs significantly from peers with similar experiences." (p. 105)*
>
> - Wolfram, Adger & Christian, 1999

For ELs, the Problem is Test Score Validity

NO BIAS

- ***Items***
  (content, novelty)
- ***Structure***
  (sequence, order, difficulty)
- ***Reliability***
  (measurement error/accuracy)
- ***Factor structure***
  (theoretical structure, relationship of variables to each other)
- ***Predictive Validity***
  (correlation with academic success or achievement)
- ***Differential Item Functioning***
  (DIF is not often found)

POTENTIAL BIAS

- ***Construct Validity***
  (nature and specificity of the intended/measured constructs)

Even when a the intended variable is measured, inferences and interpretation may not be valid if comparability in development is lacking…

- ***Interpretive Invalidity***
  (it can undermine the validity of evaluative judgments and meaning assigned to scores)

"As long as tests do not at least sample in equal degree a state of saturation [assimilation of fundamental experiences and activities] that is equal for the 'norm children' and the particular bilingual child it cannot be assumed that the test is a valid one for the child." *Sanchez, 1934*

Test Score Validity and Defensible Interpretation Requires "True Peer" Comparison

Example of Potential Construct Invalidity:

"Using these blocks, arrange them together in the correct manner so they are identical to this picture."

A test designed to measure visual processing (Gv) in ELs must avoid over-reliance on language ability (Gc) or else measurement of visual processing may be confounded with language ability.

Example of Potential Interpretive Invalidity:

"After putting a blue block on top of a purple one, put the green block on the blue one."

A test designed to measure English language ability (Gc) is valid for EL's ability *in English*, but poor performance cannot be ascribed to a potential disability unless developmental differences in English have been controlled.

Main Threats to Test Score Validity for ELLs

Acculturative Knowledge Acquisition – Not Race or Ethnicity

"When a child's general background experiences differ from those of the children on whom a test was standardized, then the use of the norms of that test as an index for evaluating that child's current performance or for predicting future performances may be inappropriate."

Salvia & Ysseldyke, 1991

Developmental Language Proficiency – Not Language Dominance

"Most studies compare the performance of students from different ethnic groups…rather than ELL and non-ELL children within those ethnic groups….A major difficulty with all of these studies is that the category Hispanic includes students from diverse cultural backgrounds with markedly different English-language skills….This reinforces the need to separate the influences of ethnicity and ELL status on observed score differences."

Lohman, Korb & Lakin, 2008

Processes and Procedures for Addressing Test Score Validity

IX. REDUCE BIAS IN TRADITIONAL TESTING PRACTICES

Exactly how is evidence-based, nondiscriminatory assessment conducted and to what extent is there any research to support the use of any of these methods in being capable of establishing sufficient validity of the obtained results?

- **Modified Methods of Evaluation**
  - *Modified and altered assessment*
- **Nonverbal Methods of Evaluation**
  - *Language reduced assessment*
- **Dominant Language Evaluation: L1**
  - *Native language assessment*
- **Dominant Language Evaluation: L2**
  - *English language assessment*

Processes and Procedures for Addressing Test Score Validity

Modified and Altered Assessment

According to the NASP Position Statement:

“monolingual school psychologists will require training in the use of interpreters in all aspects of the assessment process, as well as an awareness of the complexity of issues that may be associated with reliance on interpreters“ (p. 2).

Processes and Procedures for Addressing Test Score Validity

ISSUES IN MODIFIED METHODS OF EVALUATION

Modified and Altered Assessment:

- *use of a translator/interpreter for administration helps overcome the language barrier but is also a violation of standardization and undermines score validity, even when the interpreter is highly trained and experienced; tests are not usually normed in this manner*
- *in efforts to help the examinee perform to the best of his/her ability, any process involving "testing the limits" where there is alteration or modification of test items or content, mediation of task concepts prior to administration, repetition of instructions, acceptance of responses in either languages, or elimination/modification of time constraints, etc., violates standardization even when "permitted" by the test publisher except in cases where separate norms for such altered administration are provided*
- *any alteration of the testing process violates standardization and effectively invalidates the scores which precludes interpretation or the assignment of meaning by undermining the psychometric properties of the test*
- *alterations or modifications are perhaps most useful in deriving qualitative information—observing behavior, evaluating learning propensity, evaluating developmental capabilities, analyzing errors, etc.*
- *a recommended procedure would be to administer tests in a standardized manner first, which will potentially allow for later interpretation, and then consider any modifications or alterations that will further inform the referral questions*
- *because the violation of the standardized test protocol introduces error into the testing process,* ***it cannot be determined to what extent the procedures aided or hindered performance and thus the results cannot be defended as valid***

Processes and Procedures for Addressing Test Score Validity

Nonverbal Assessment

According to the NASP Position Statement:

"the use of 'nonverbal" tools or native language instruments are not automatic guarantees of reliable and valid data. Nonverbal tests rely on some form of effective communication between examiner and examinee, and may be as culturally loaded as verbal tests, thus limiting the validity of evaluation results." (p. 2).

Processes and Procedures for Addressing Test Score Validity

ISSUES IN NONVERBAL METHODS OF EVALUATION

Language Reduced Assessment:

- *"nonverbal testing:" use of language-reduced (or 'nonverbal') tests are helpful in overcoming the language obstacle, however:*
- *it is impossible to administer a test without some type of communication occurring between examinee and examiner, this is the purpose of gestures/pantomime*
- *some tests remain very culturally embedded—they do not become culture-free simply because language is not required for responding*
- *construct underrepresentation is common, especially on tests that measure fluid reasoning (Gf), and when viewed within the context of CHC theory, some batteries measure a narrower range of broad cognitive abilities/processes, particularly those related to verbal academic skills such as reading and writing (e.g., Ga and Gc) and mathematics (Gq)*
- *all nonverbal tests are subject to the same problems with norms and cultural content as verbal tests—that is, they do not control for differences in acculturation and language proficiency which may still affect performance, albeit less than with verbal tests*
- *language reduced tests are helpful in evaluation of diverse individuals and may provide better estimates of true functioning in certain areas,* ***but they are not a whole or completely satisfactory solution with respect to fairness and provide no mechanism for establishing whether the obtained test results are valid or not***

Processes and Procedures for Addressing Test Score Validity

Native Language Assessment (L1)

According to the NASP Position Statement:

"NASP supports the rights of bilingual students who are referred for a psychoeducational evaluation to be assessed in their native languages when such evaluation will provide the most useful data to inform interventions…Furthermore, the norms for native language tests may not represent the types of ELLs typically found in U.S. schools, and very limited research exists on how U.S. bilingual students perform on tests in their native language as opposed to English." (p. 2).

Processes and Procedures for Addressing Test Score Validity

ISSUES IN DOMINANT LANGUAGE EVALUATION: Native language

Native Language Assessment (L1):

- *generally refers to the assessment of bilinguals by a bilingual psychologist who has determined that the examinee is more proficient ("dominant") in their native language than in English*
- *being "dominant" in the native language does not imply age-appropriate development in that language or that formal instruction has been in the native language or that both the development and formal instruction have remained uninterrupted in that language*
- *although the bilingual psychologist is able to conduct assessment activities in the native language, this option is not directly available to the monolingual psychologist*
- *native language assessment is a relatively new idea and an unexplored research area so there is very little empirical support to guide appropriate activities or upon which to base standards of practice or evaluated test performance*
- *whether a test evaluates only in the native language or some combination of the native language and English (i.e., presumably "bilingual"), the norm samples may not provide adequate representation or any at all on the critical variables (language proficiency and acculturative experiences)—bilinguals in the U.S. are not the same as monolinguals elsewhere*
- ***without a research base, there is no way to evaluate the validity of the obtained test results*** *and any subsequent interpretations would be specious and amount to no more than a guess*

ELL Test Performance: Esparza Brown Study

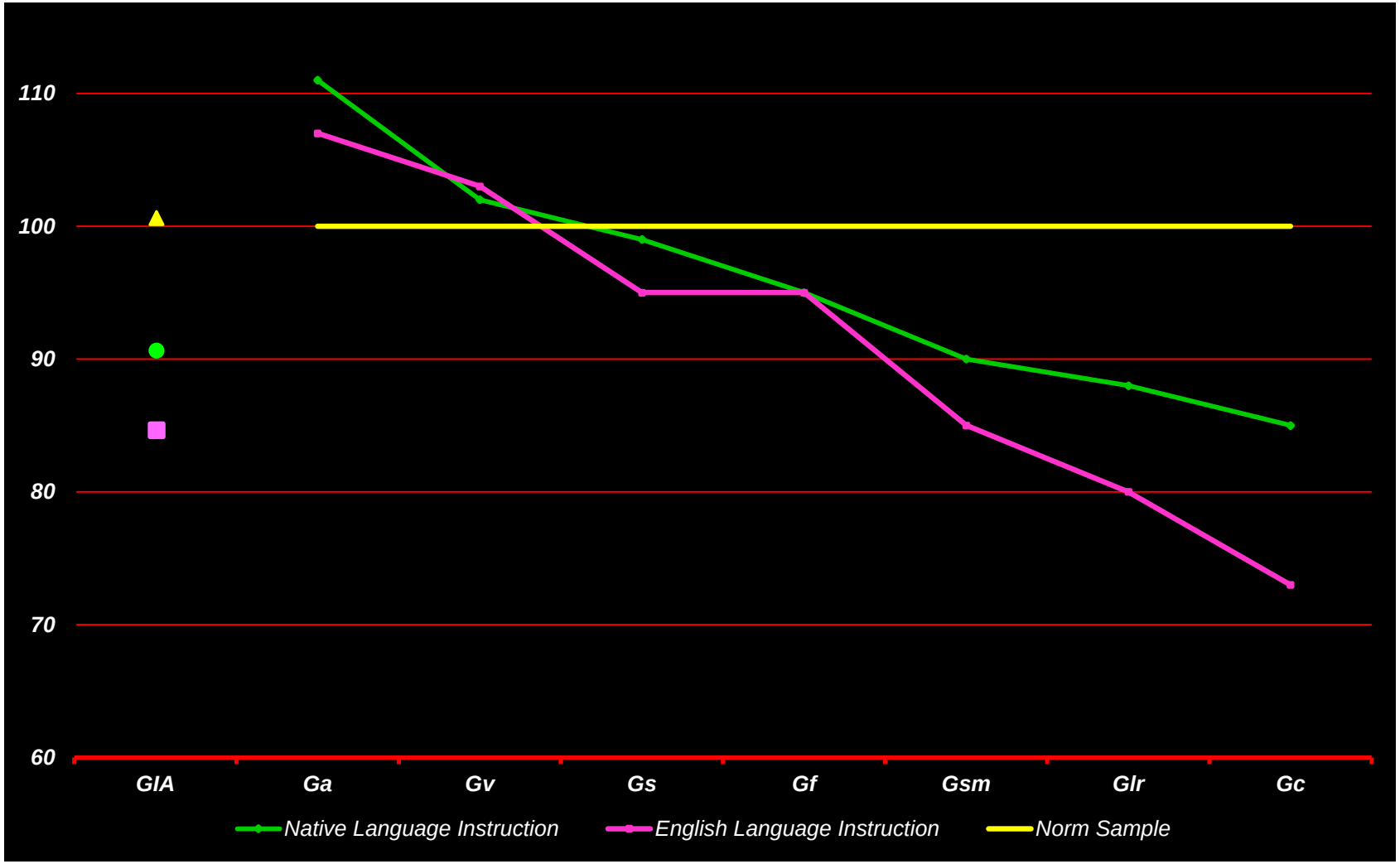
Comparison of Order of Means for WJ III and Bateria III Classifications*

WJ III Classifications		Bateria III Classifications (NLD)		Bateria III Classifications (ELD)	
Mean	**Subtest**	**Mean**	**Subtest**	**Mean**	**Subtest**
98	Gv – Visual Processing	111	Ga – Auditory Processing	107	Ga – Auditory Processing
95	Gs – Processing Speed	102	Gv – Visual Processing	103	Gv – Visual Processing
95	Gsm – Short Term Memory	99	Gs – Processing Speed	95	Gs – Processing Speed
92	Gf – Fluid Reasoning	95	Gf – Fluid Reasoning	95	Gf – Fluid Reasoning
89	Ga – Auditory Processing	90	Glr – Long Term Memory	82	Gsm – Short Term Memory
89	Glr – Long Term Memory	88	Gsm – Short Term Memory	77	Glr – Long Term Memory
85	Gc – Crystallized Knowledge	85	Gc – Crystallized Knowledge	73	Gc – Crystallized Knowledge

**Source: Esparza Brown, J. (2008). The use and interpretation of the Bateria III with U.S. Bilinguals. Unpublished dissertation, Portland State University, Portland, OR.*

ELL Test Performance: Esparza Brown Study

Comparison of Bateria III Cluster Means for ELL's by Language of Instruction



*Source: Esparza Brown, J. (2008). The use and interpretation of the Bateria III with U.S. Bilinguals. Unpublished dissertation, Portland State University, Portland, OR.

Processes and Procedures for Addressing Test Score Validity

English Language Assessment (L2)

According to the NASP Position Statement:

"monolingual, English-speaking school psychologists will likely conduct the vast majority of evaluations with bilingual students. Therefore, proper training in the requisite knowledge and skills for culturally and linguistically responsive assessment is necessary for ***all*** school psychologists." (p. 2; emphasis added).

Processes and Procedures for Addressing Test Score Validity

ISSUES IN DOMINANT LANGUAGE EVALUATION: English

English Language Assessment (L2):

- *generally refers to the assessment of bilinguals by a monolingual psychologist who had determined that the examinee is more proficient ("dominant") in English than in their native language or without regard to the native language at all*
- *being "dominant" in the native language does not imply age-appropriate development in that language or that formal instruction has been in the native language or that both the development and formal instruction have remained uninterrupted in that language*
- *does not require that the evaluator speak the language of the child but does require competency, training and knowledge, in nondiscriminatory assessment including the manner in which cultural and linguistic factors affect test performance*
- *evaluation conducted in English is a very old idea and a well explored research area so there is a great deal of empirical support to guide appropriate activities and upon which to base standards of practice and evaluate test performance*
- *the greatest concern when testing in English is that the norm samples of the tests may not provide adequate representation or any at all on the critical variables (language proficiency and acculturative experiences)—dominant English speaking ELLs in the U.S. are not the same as monolingual English speakers in the U.S.*
- *with an extensive research base, **the validity of the obtained test results may be evaluated** (e.g., via use of the Culture-Language Interpretive Matrix) and would permit defensible interpretation and assignment of meaning to the results*

Current Approaches Fail to Establish Test Score Validity

Evaluation Issues and Methods	Norm sample representative of bilingual development	Measures a wider range of school-related abilities	Does not require the evaluator to be bilingual	Adheres to the test's standardized protocol	Substantial research base on bilingual performance	Sufficient to identify or diagnosis disability	Accounts for variation in bilingual development	Most likely to yield reliable and valid data and information	Provides extensive data regarding development
Modified or Altered Assessment	✗	✓	✓	✗	✗	✗	✗	✗	✗
Language Reduced Assessment	✗	✗	✓	✓	✗	✗	✗	✗	✗
Dominant Language Assessment in L1: native only	✗	✓	✗	✓	✗	✗	✗	✗	✗
Dominant Language Assessment in L2: English only	✗	✓	✓	✓	✓	✗	✗	✗	✗

All approaches are limited in some manner when addressing test score validity and none are sufficient to diagnosis a disability, account for variation in bilingual development, represent a form or manner that automatically yields reliable and valid results, and do not provide extensive data regarding cognitive and school-based learning and development.

Test Score Validity and Defensible Interpretation Requires "True Peer" Comparison

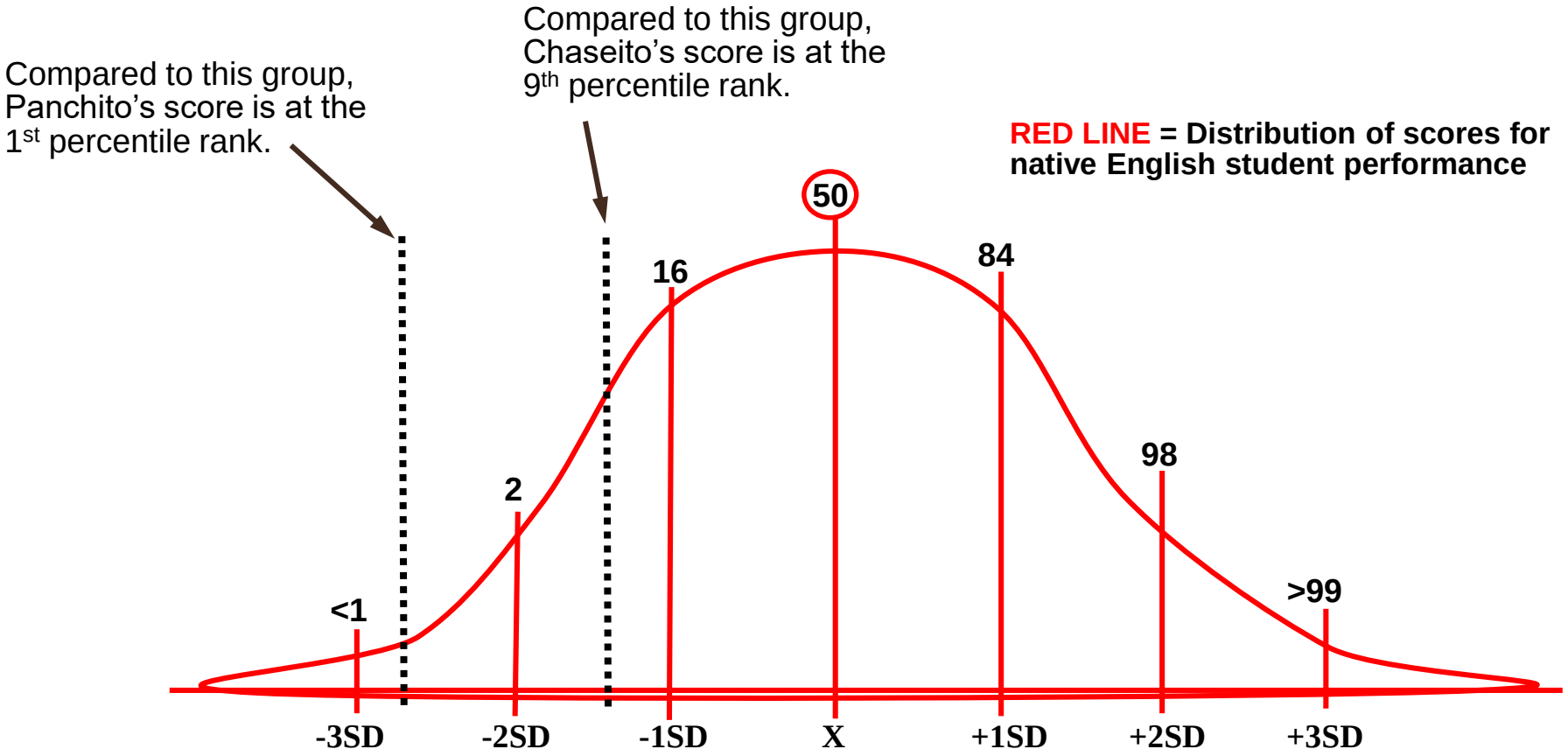
For native English speakers, growth of language-related abilities are tied closely to age because the process of learning a language begins at birth and is fostered by formal schooling. Thus, age-based norms effectively control for variation in development and provide an appropriate basis for comparison. However, this is not true for English learners who may begin learning English at various points after birth and who may receive vastly different types of formal education from each other.

Development Varies by Exposure to English – Not dominance

"It is unlikely that a second-grade English learner at the early intermediate phase of language development is going to have the same achievement profile as the native English-speaking classmate sitting next to her. The norms established to measure fluency, for instance, are not able to account for the language development differences between the two girls. A second analysis of the student's progress compared to linguistically similar students is warranted." (p. 40)

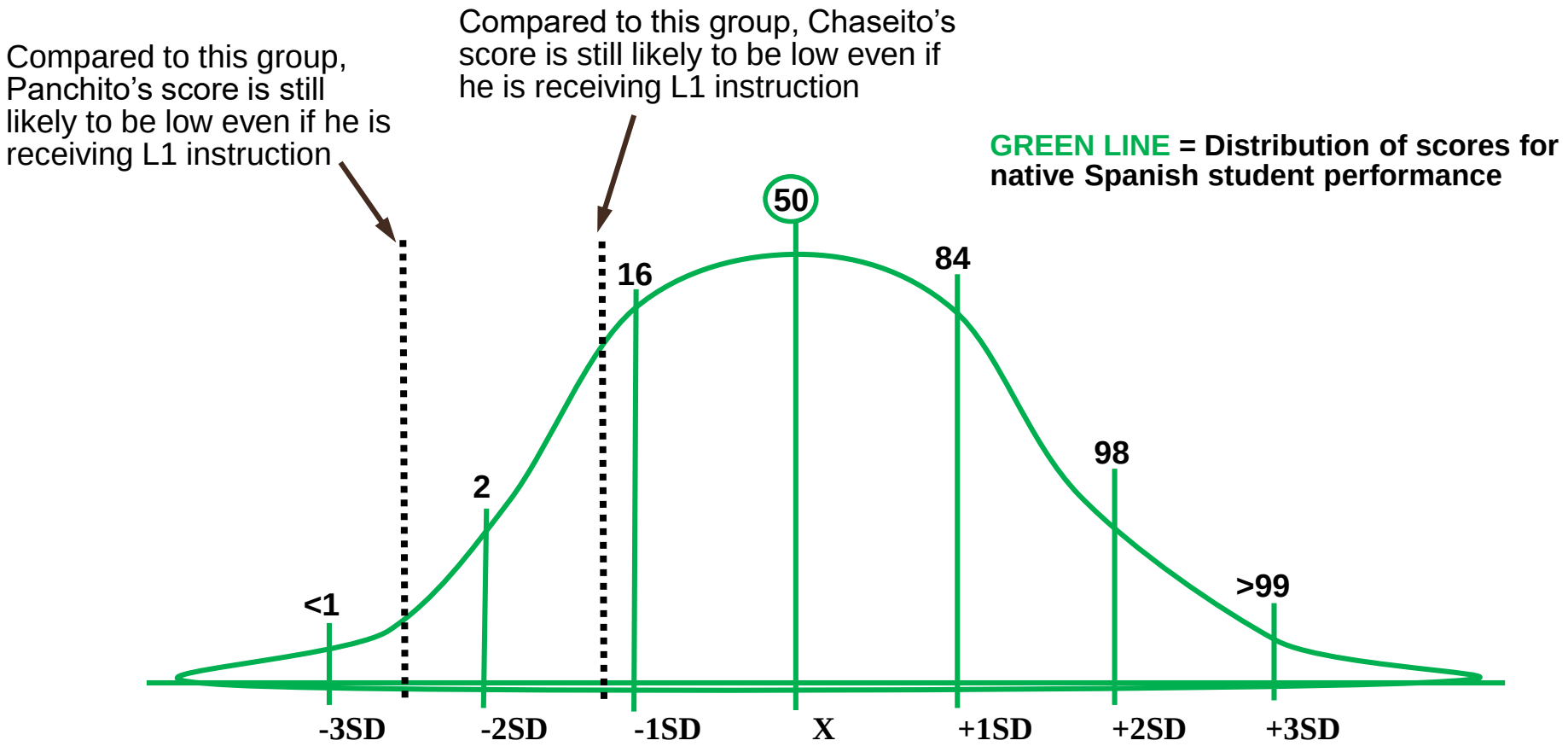
- Fisher & Fry, 2012

The validity of an interpretation regarding disability requires an unbiased standard for comparison.



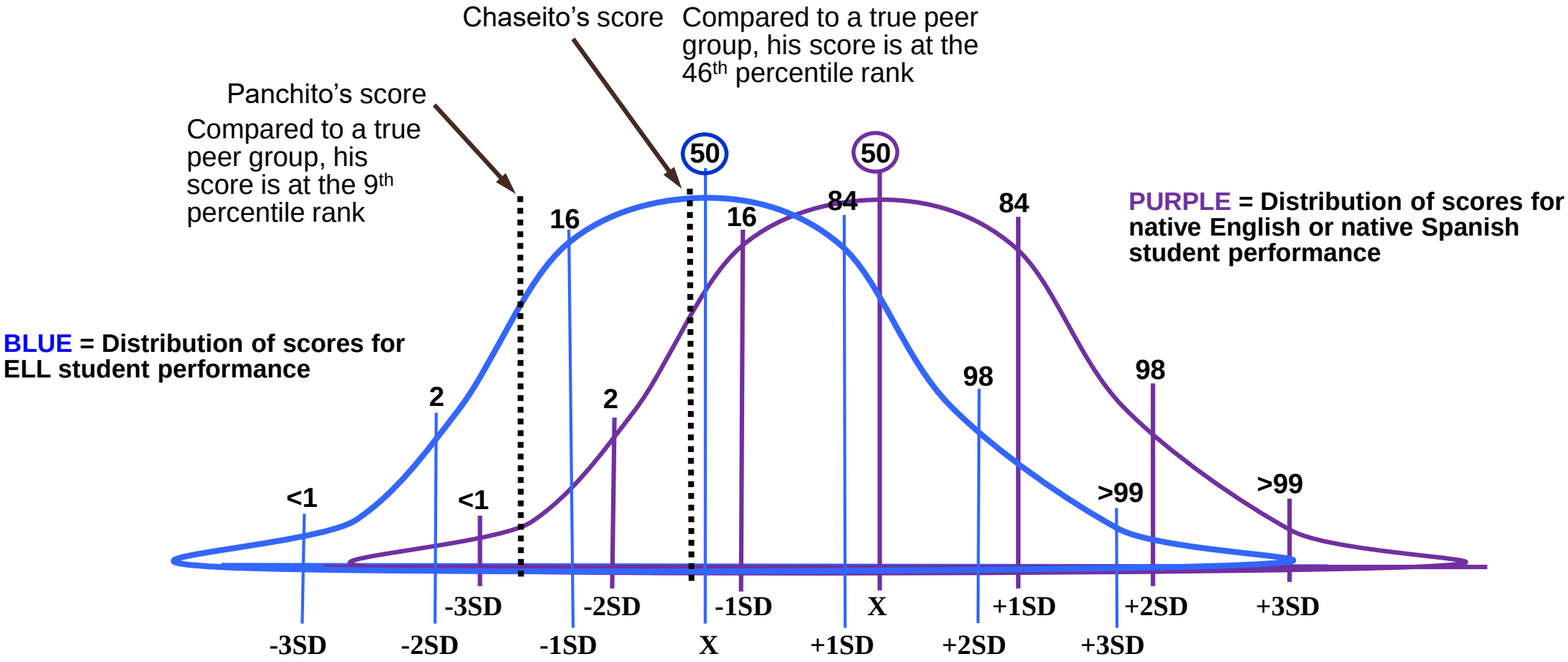
Using an inappropriate comparison group makes it appear incorrectly that both Chaseito and Panchito may have some type of disability.

The validity of an interpretation regarding disability requires an unbiased standard for comparison.



Use of a native-language group remains an inappropriate comparison and continues to make it appear incorrectly that both Chaseito and Panchito have some type of disability.

The validity of an interpretation regarding disability requires an unbiased standard for comparison.



Use of a "true peer" group provides a non-discriminatory comparison and suggests that Chaseito's performance is average and that only Panchito might have some type of disability.

The validity of an interpretation regarding disability requires an unbiased standard for comparison.

Whatever method or approach may be employed in evaluation of ELL's, the fundamental obstacle to nondiscriminatory interpretation rests on the degree to which the examiner is able to defend claims of test score construct validity that is being used to support diagnostic conclusions. This idea is captured by and commonly referred to as a question of:

"DIFFERENCE vs. DISORDER?"

Simply absolving oneself from responsibility of establishing test score validity, for example via wording such as, "all scores should be interpreted with extreme caution" does not in any way provide a defensible argument regarding the validity of obtained test results and does not permit valid diagnostic inferences or conclusions to be drawn from them.

The only manner in which test score validity can be evaluated or established to a degree that permits valid and defensible diagnostic inferences with ELL's is to use a comparison standard that represents "true peers."

Evidence-Based Assessment

According to the APA Task Force on Evidence-based practice in psychology (EBPP), evidence-based practice is defined as:

> *"the integration of the best available research with clinical expertise in the context of patient characteristics, culture, and preferences (p. 273)*

Evidence-based practice within the context of psychoeducational evaluation has never gone much beyond an over-reliance on the validity of standardized tests. But without inherently fair norm samples, the only recourse for individual practitioners is to apply research on the use of standardized tests with English learners. This becomes, in effect, evidence-based assessment.

Source: American Psychological Association (2006). Evidence-Based Practice in Psychology, American Psychologist, pp. 271-285.

Summary of Research on the Test Performance of English Language Learners

Research conducted over the past 100 years on ELLs who are non-disabled, of average ability, possess moderate to high proficiency in English, and tested in English, has resulted in two robust and ubiquitous findings:

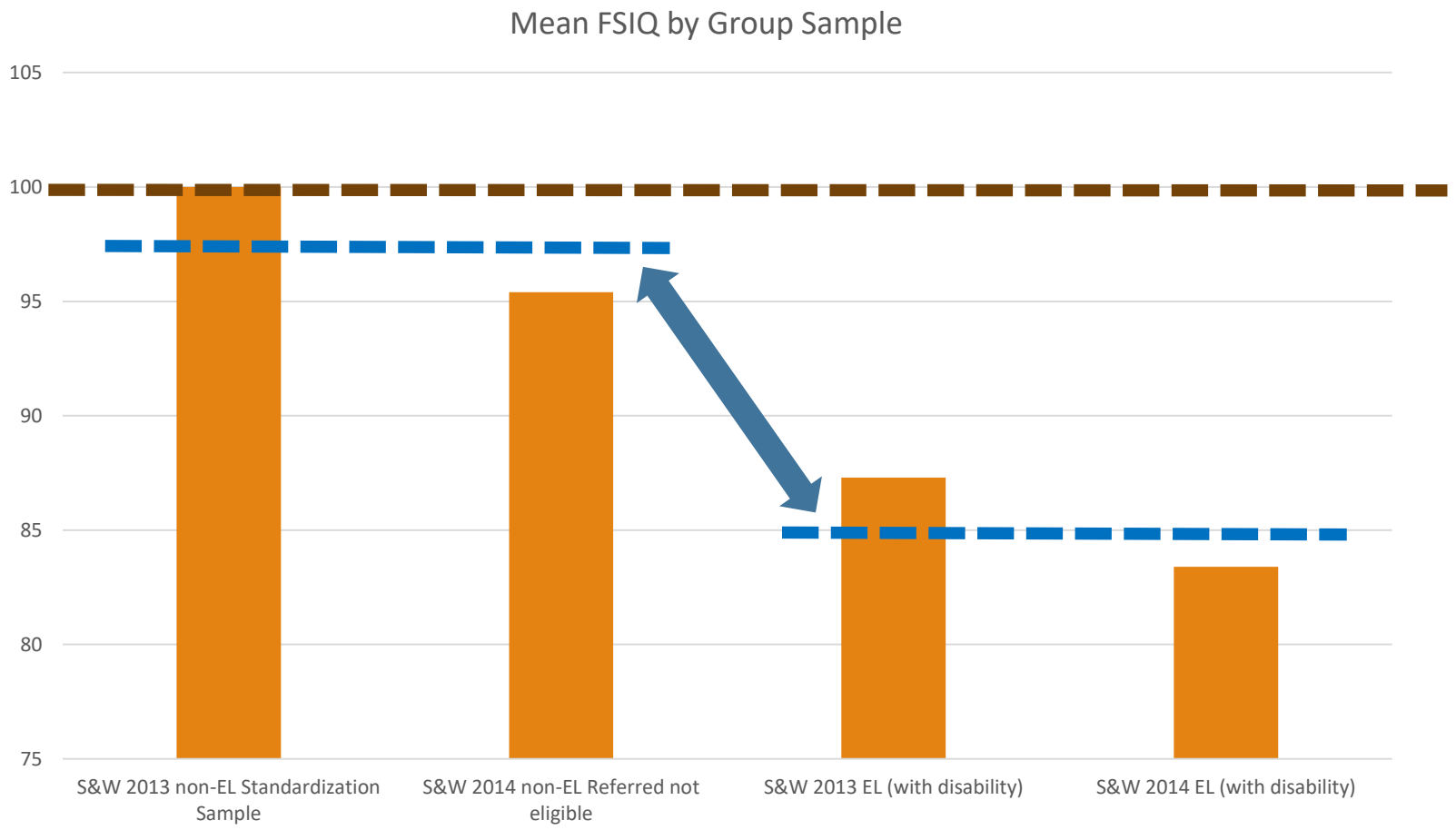
1. *Native English speakers perform better than English learners at the broad ability level (e.g., FSIQ) on standardized, norm-referenced tests of intelligence and general cognitive ability.*

2. *English learners tend to perform significantly better on nonverbal type tests than they do on verbal tests (e.g., PIQ vs. VIQ).*

So what explains these findings? Early explanations relied on genetic differences attributed to race even when data strongly indicated that the test performance of ELLs was moderated by the degree to which a given test relied on or required age- or grade-expected development in English and the acquisition of incidental acculturative knowledge.

Research Foundations for EL Evaluation

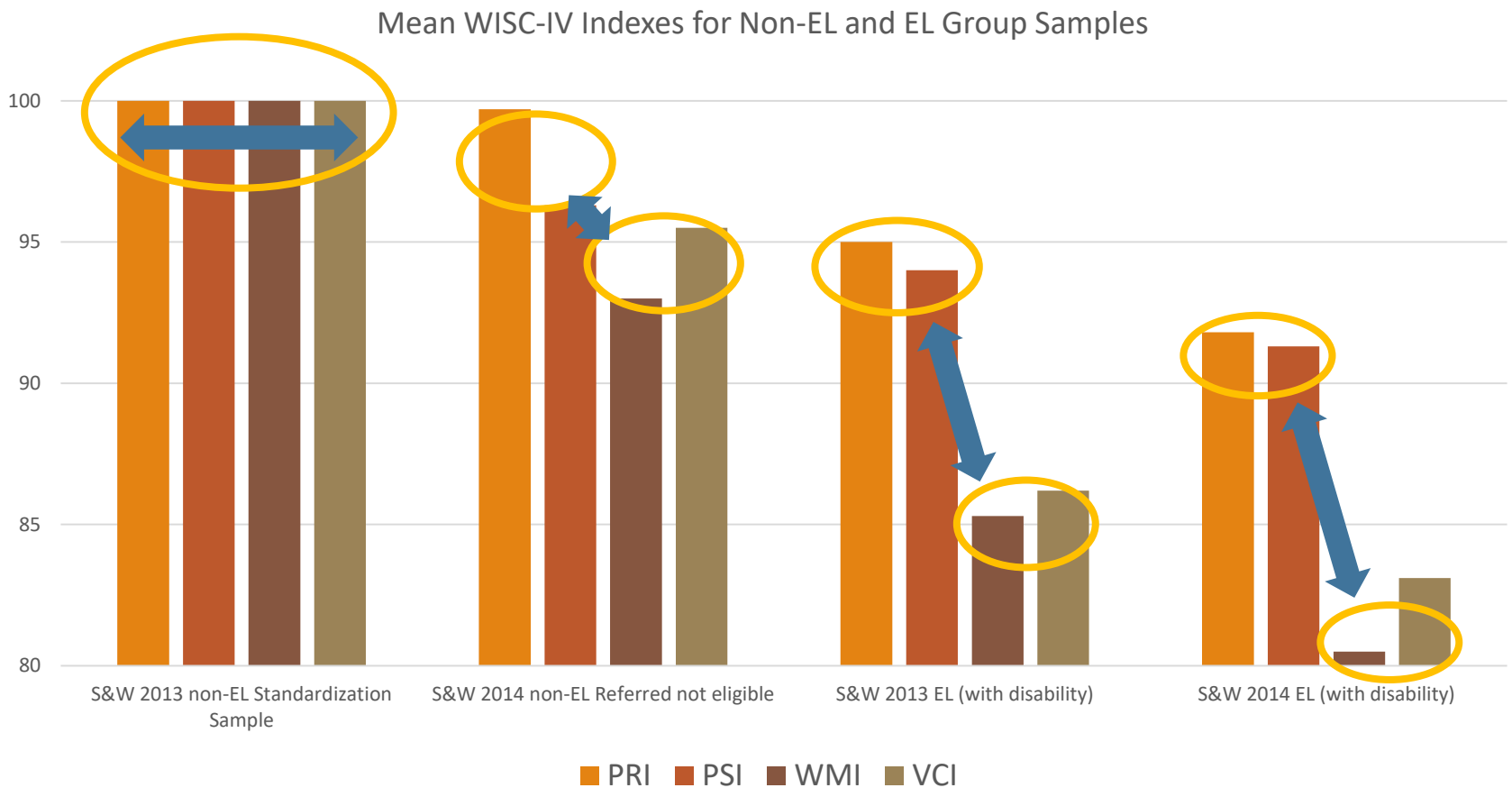
Principle 1: Native English Speakers perform better than ELs at the broad ability level.



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 EL Evaluation

Principle 2: ELs perform better on nonverbal tests than verbal tests



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 ELL Evaluation

Historical and contemporary research has tended to ignore the fact that ELLs do not perform at the same level on ALL nonverbal tests any more than they perform at the same level on ALL verbal tests.

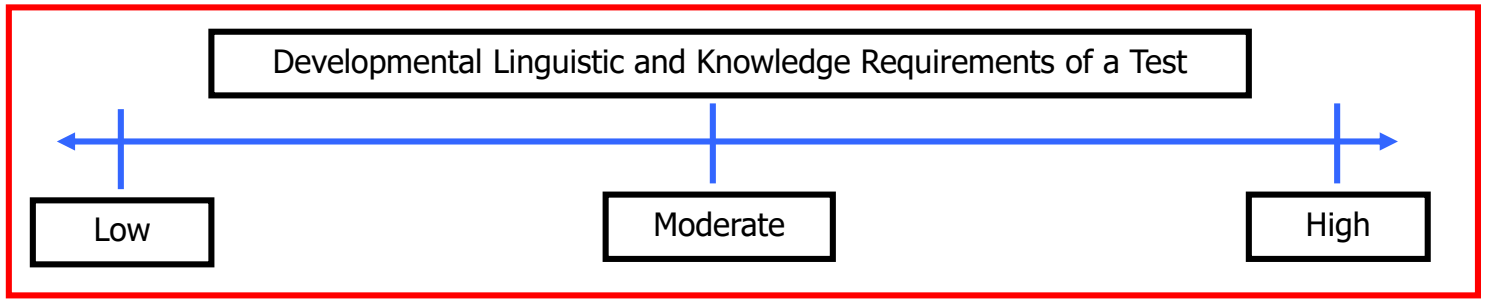
Instead, it appears that test performance of ELLs is not a dichotomy but rather a continuum formed by a linear, not dichotomous, attenuation of performance.

This means, a third principle is evident in the body of research on ELLs but has not been well understood or utilized in understanding test performance:

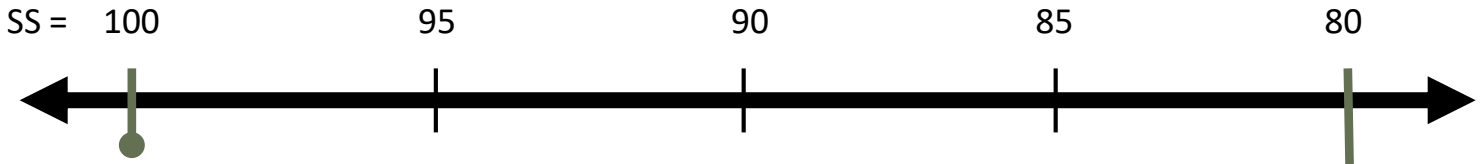
3. *Test performance of ELLs is moderated by the degree to which a given test relies on or requires age- or grade-expected English language development and the acquisition of incidental acculturative knowledge.*

Research Foundations for ELL Evaluation

EL test performance is a linear, continuous pattern, not a dichotomy.



The more a test requires age-based developmental language proficiency and acculturative knowledge, the more the effect on test performance.

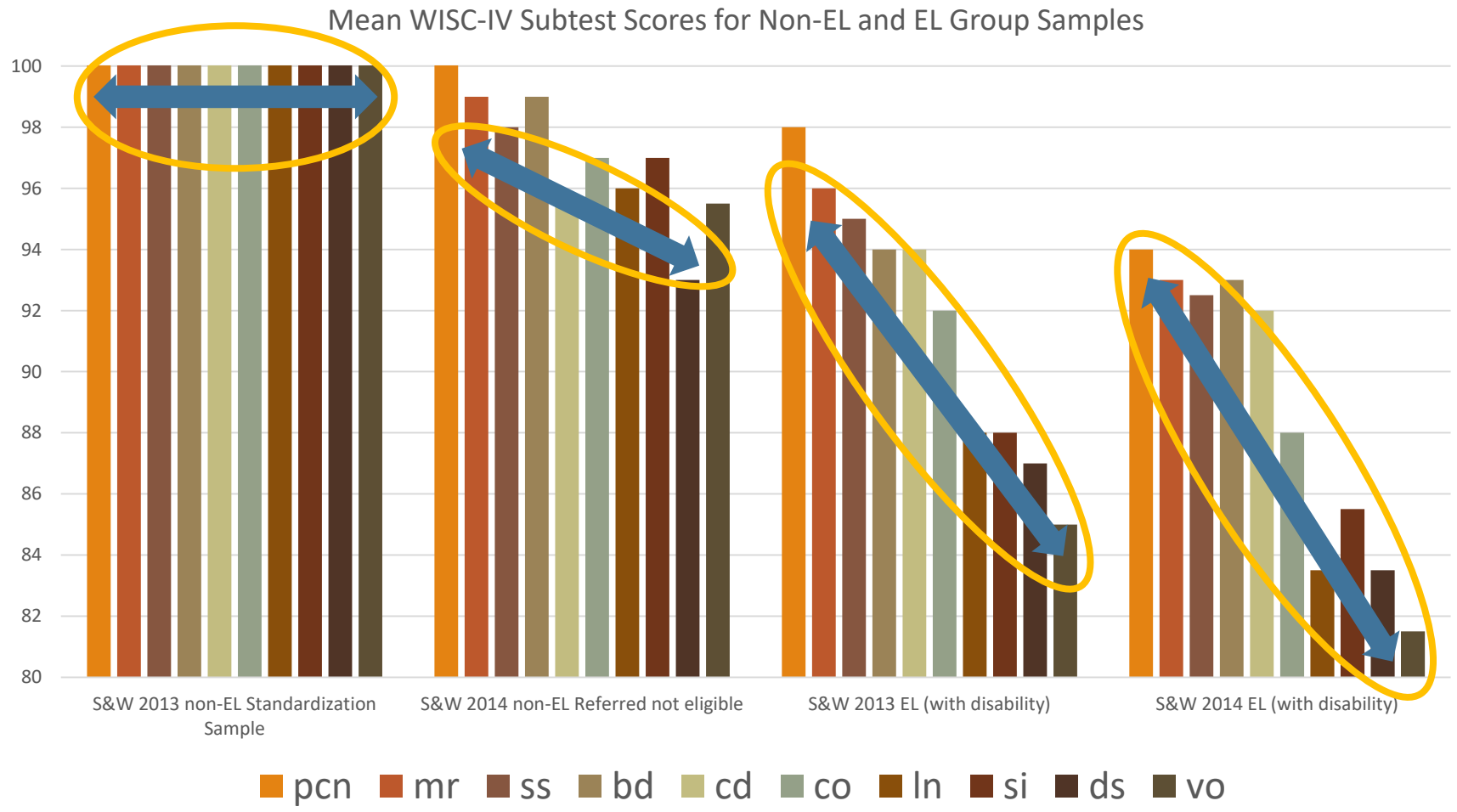


For ELs, tests that require little or no level of age-based acquisition of language and cultural knowledge yield scores at or close to the normative mean.

For ELs, tests that require full or high levels of age-based acquisition of language and cultural knowledge yield scores much lower than the normative mean.

Research Foundations for ELL Evaluation

Principle 3: ELL performance is moderated by linguistic/acculturative variables



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 ELL Evaluation

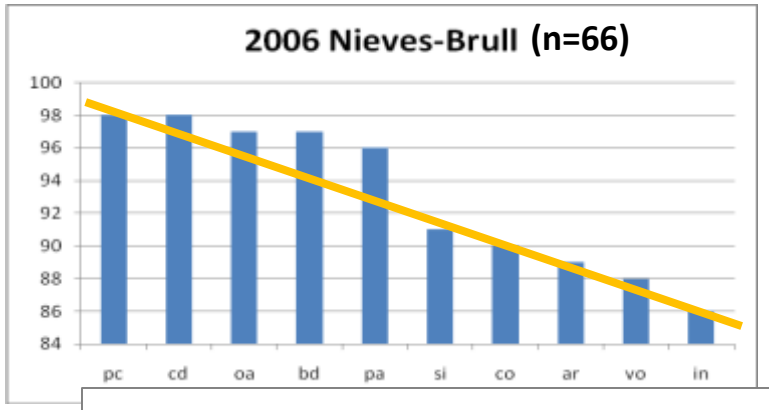
Principle 3: ELL performance is moderated by linguistic/acculturative variables

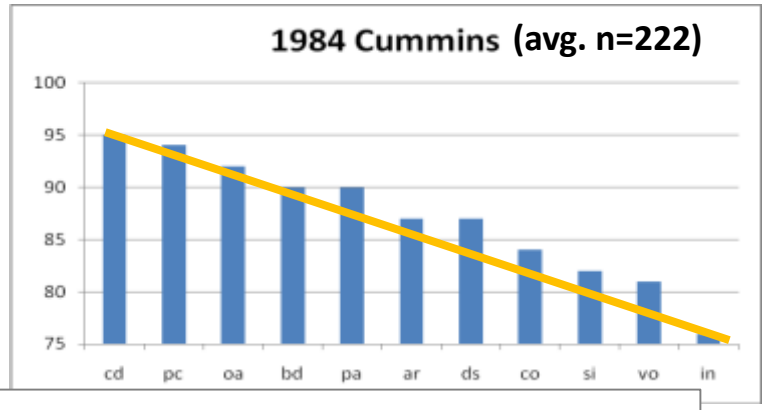
	Hispanic Group (Mercer) (1972)	Hispanic Group (Vukovich & Figueroa) (1982)	ESL Group (Cummins) (1982)	Bilingual Group (Nieves-Brull) (2006)
Subtest Name	**Mean SS**	**Mean SS**	**Mean SS**	**Mean SS**
Information	7.5	7.8	5.1	7.2
Vocabulary	8.0	8.3	6.1	7.5
Similarities	7.6	8.8	6.4	8.2
Comprehension	7.8	9.0	6.7	8.0
Digit Span	8.3	8.5	7.3	*
Arithmetic	8.7	9.4	7.4	7.8
Picture Arrangement	9.0	10.3	8.0	9.2
Block Design	9.5	10.8	8.0	9.4
Object Assembly	9.6	10.7	8.4	9.3
Picture Completion	9.7	9.9	8.7	9.5
Coding	9.6	10.9	8.9	9.6

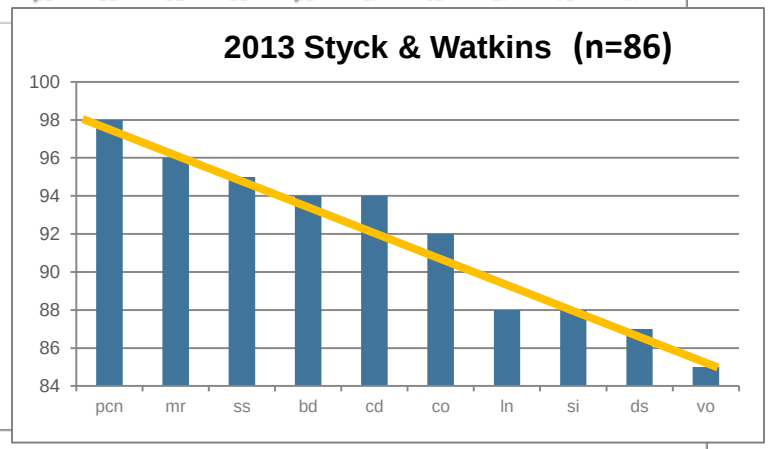
**Data for this subtest were not reported in the study.*

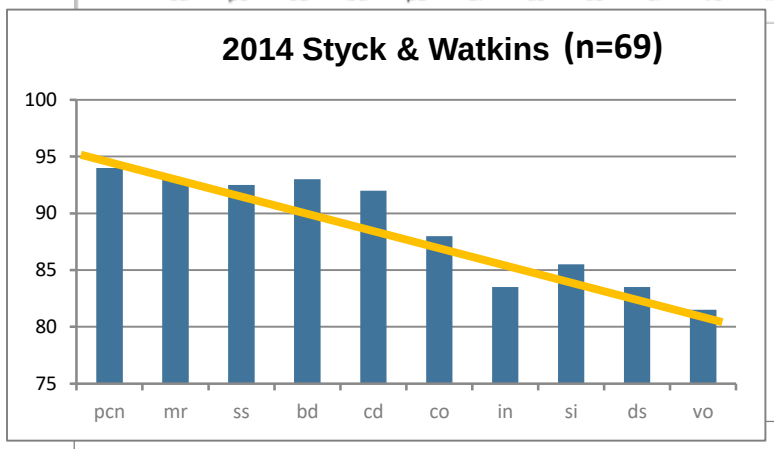
Research Foundations for ELL Evaluation

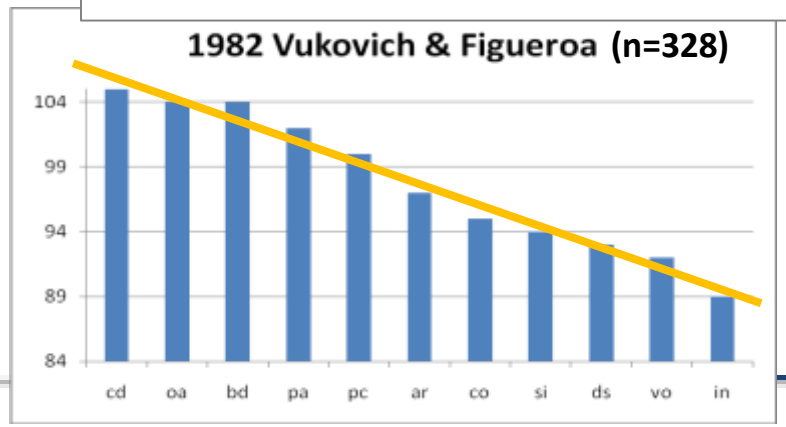
Principle 3: ELL performance is moderated by linguistic/acculturative variables

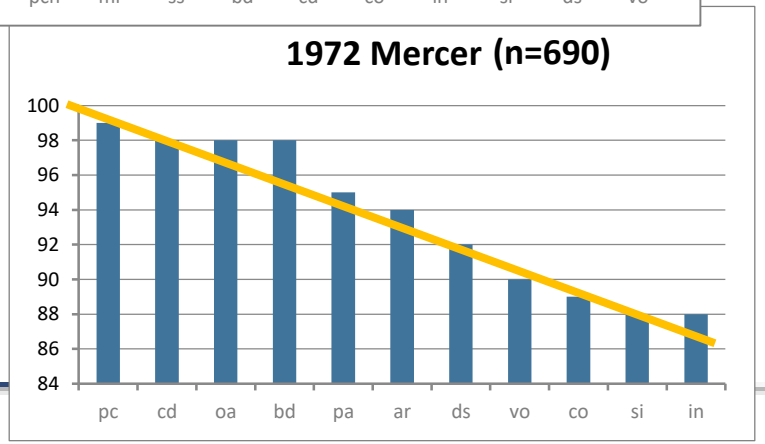












Research Foundations for EL Evaluation

Principle 3: EL performance is moderated by linguistic/acculturative variables

Highest Language Demands

Lowest Language Demands

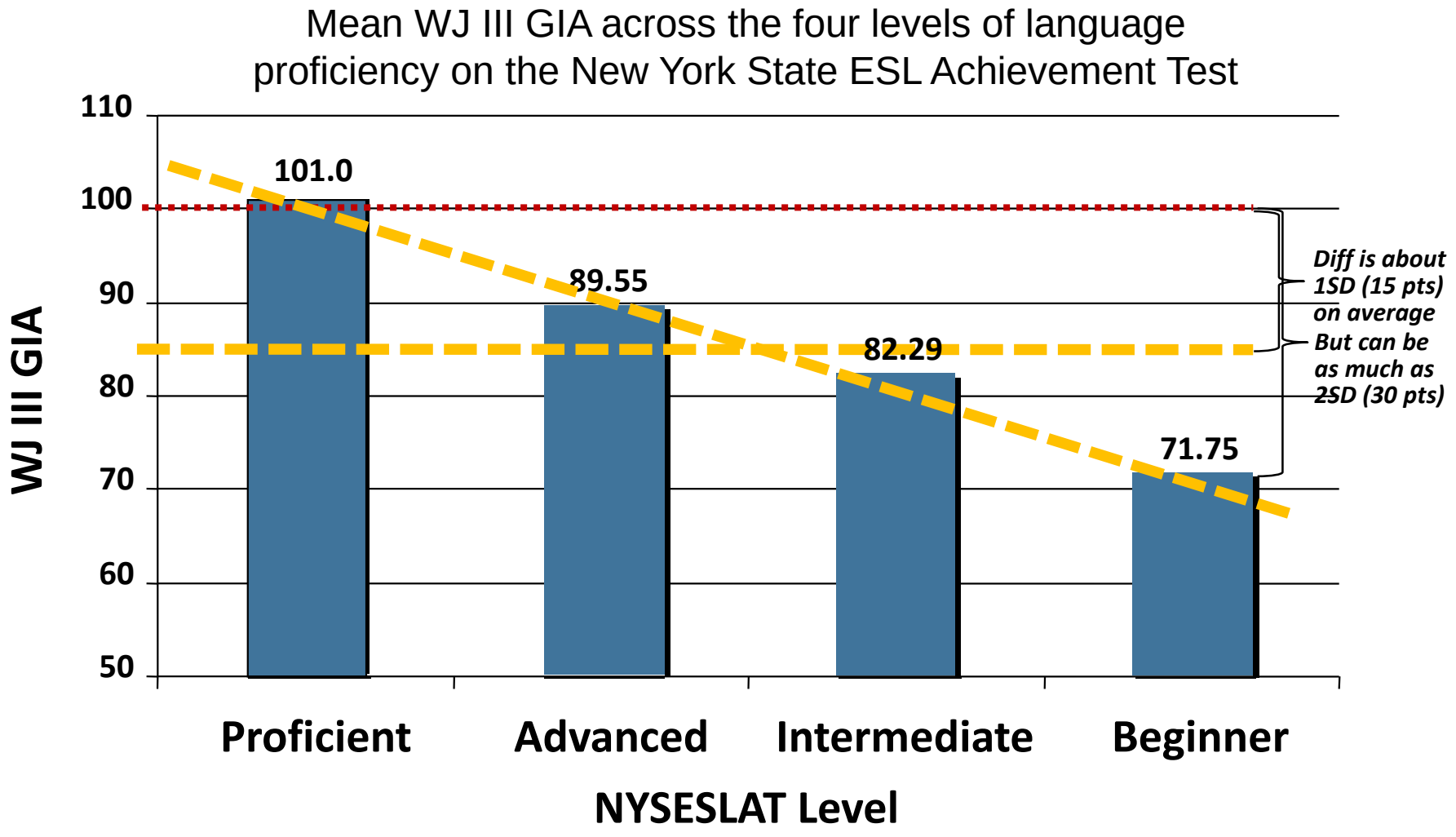
Table 3. Variance Explained by Exogenous Variables (Individual Test Performance) by Age Group.

		Variance explained		
	Individual test	7-10	11-14	15-18
Tier 5	Verbal Comprehension	.79[c]	.86[c]	.81[c]
	General Information	.71[c]	.85[c]	.86[c]
Tier 4	Concept Formation	.67[c]	.71[c]	.67[c]
	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]
Tier 3	Sound Blending	.25[b]	.32[b]	.35[b]
	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]
Tier 2	Incomplete Words	.13[b]	.31[b]	.23[b]
	Visual Matching	.13[b]	.15[b]	.16[b]
	Decision Speed	.12[b]	.15[b]	.19[b]
	Auditory Attention	.10[b]	.20[b]	.15[b]
Tier 1	Spatial Relations	.08[a]	.16[b]	.16[b]
	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 EL Evaluation

Principle 3: EL performance is moderated by linguistic/acculturative variables

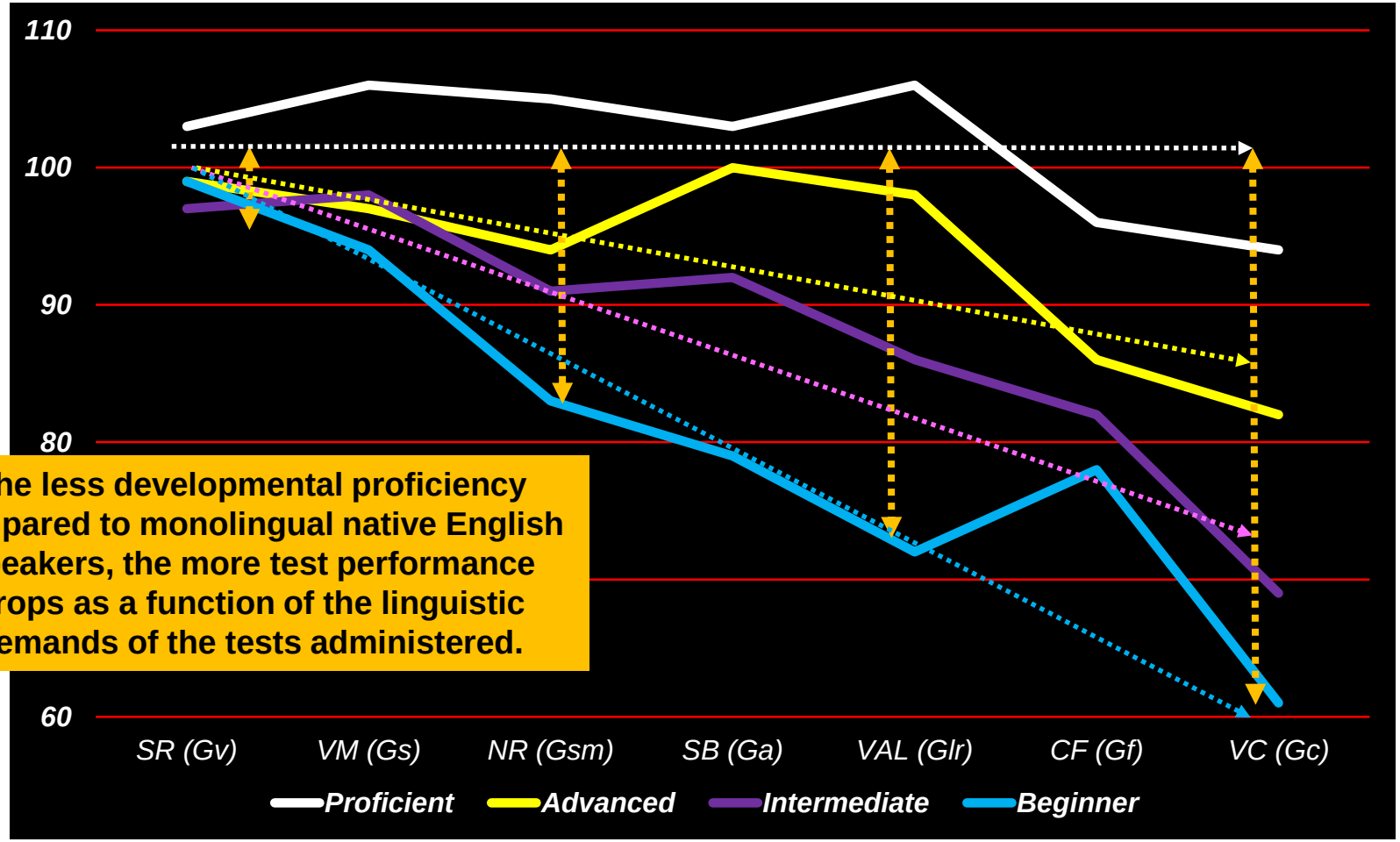


Source: Sotelo-Dynega, M., Ortiz, S.O., Flanagan, D.P., Chaplin, W. (2013).

Research Foundations for ELL Evaluation

Principle 3: ELL performance is moderated by linguistic/acculturative variables

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



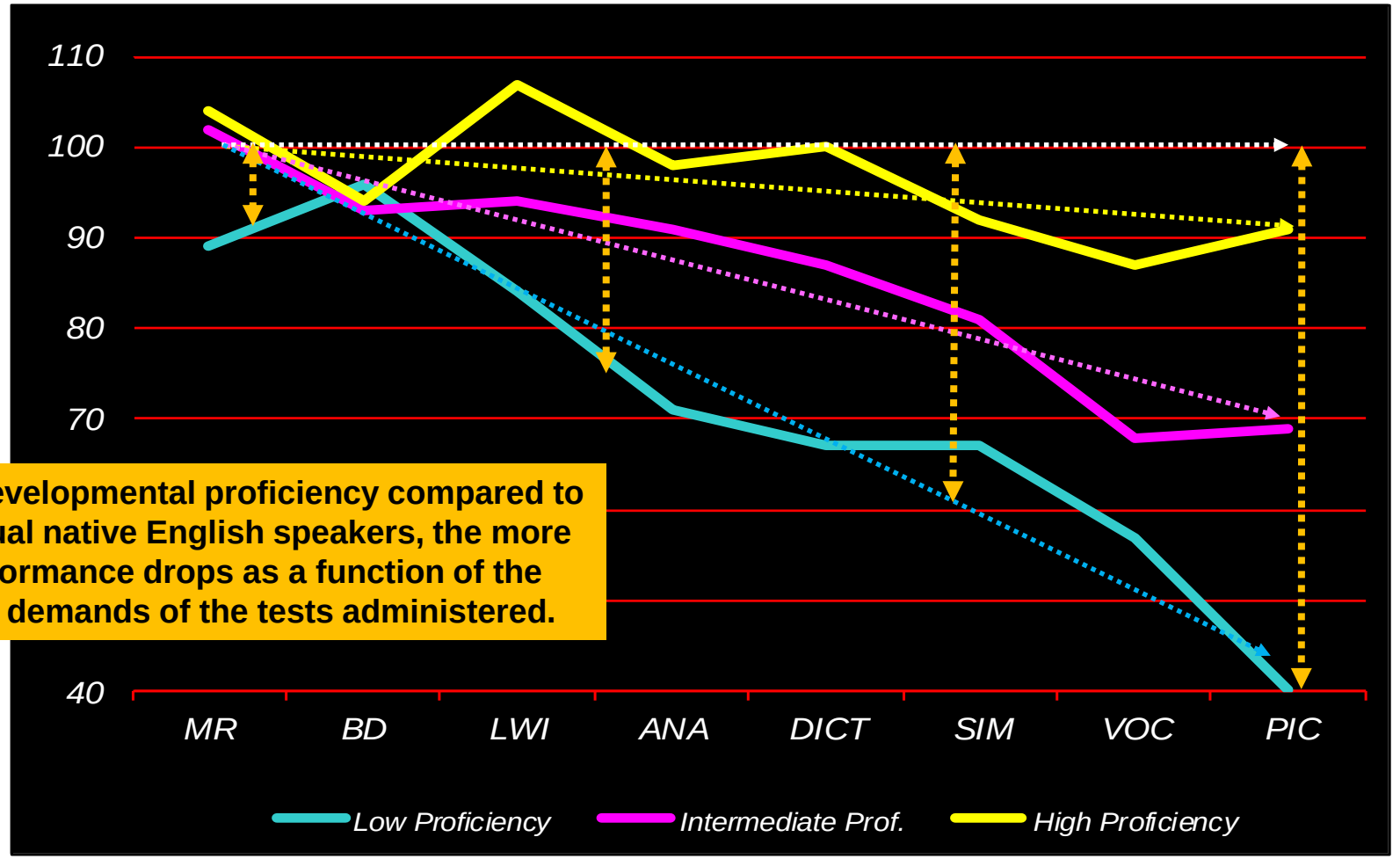
The less developmental proficiency compared to monolingual native English speakers, the more test performance drops as a function of the linguistic demands of the tests administered.

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 ELL Evaluation

Principle 3: ELL 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.

Summary of the Foundational Research Principles of the Culture-Language Interpretive Matrix

Principle 1: EL and non-EL's perform differently at the broad ability level on tests of cognitive ability.

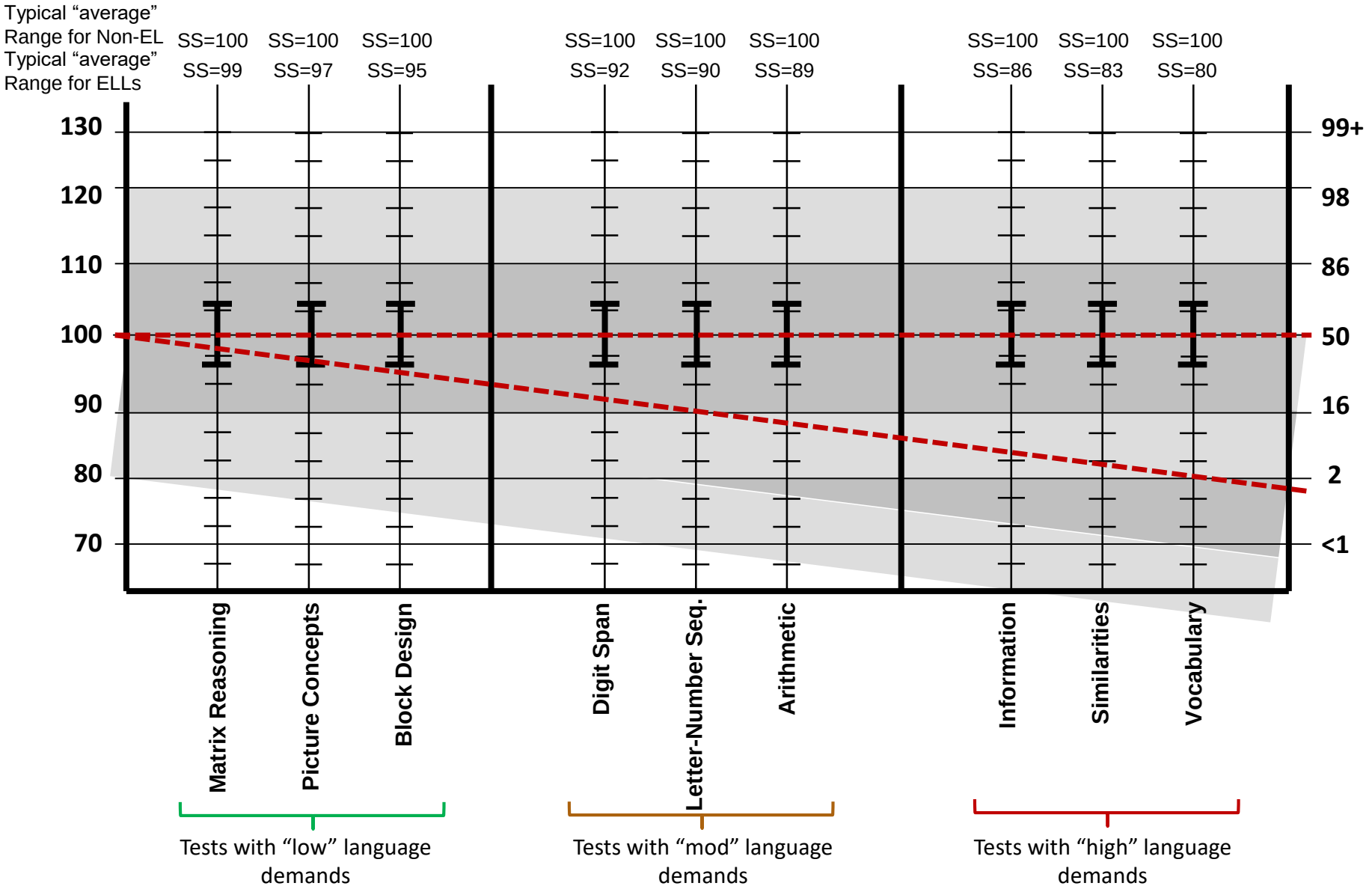
Principle 2: ELs perform better on nonverbal tests than they do on verbal tests.

Principle 3: EL performance on both verbal and nonverbal tests is moderated by linguistic and acculturative variables.

Because the basic research principles underlying the C-LIM are well supported, it means that use of the C-LIM is valid and renders it an example of evidence-based practice.

- This does not mean, however, that it cannot be improved. Productive research on EL test performance can assist in making any necessary "adjustments" to the order of the means as arranged in the C-LIM.
- Likewise, as new tests come out, new research is needed to determine the relative level of EL performance as compared to other tests with established values of expected average performance.
- Ultimately, only research that focuses on stratifying samples by relevant variables such as language proficiency, length and type of English and native language instruction, and developmental issues related to age and grade of first exposure to English, will serve useful in furthering knowledge in this area and assist in establishing appropriate expectations of test performance for specific populations of ELs.

Fairness in Determining "Average" Performance



The Culture-Language Interpretive Matrix (C-LIM)

Important Facts for Use and Practice

The C-LIM is not a test, scale, measure, or mechanism for making diagnoses. It is a visual representation of current and previous research on the test performance of English learners arranged by mean values to permit examination of the combined influence of acculturative knowledge acquisition and limited English proficiency and its impact on test score validity.

The C-LIM is not a language proficiency measure and will not distinguish native English speakers from English learners with high, native-like English proficiency and is not designed to determine if someone is or is not an English learner. Moreover, the C-LIM is not for use with individuals who are native English speakers.

The C-LIM is not designed or intended for diagnosing any particular disability but rather as a tool to assist clinician's in making decisions regarding whether ability test scores should be viewed as indications of actual disability or rather a reflection of differences in language proficiency and acculturative knowledge acquisition.

The primary purpose of the C-LIM is to assist evaluators in ruling out cultural and linguistic influences as exclusionary factors that may have undermined the validity of test scores, particularly in evaluations of SLD or other cognitive-based disorders. Being able to make this determination is the primary and main hurdle in evaluation of ELLs and the C-LIM's purpose is to provide an evidence-based method that assists clinician's regarding interpretation of test score data in a nondiscriminatory manner.

Evaluation Resources for Evaluation of English Learners

General C-LIM web site with full file listing: http://facpub.stjohns.edu/~ortizs/CLIM/

Culture-Language Interpretive Matrix – Non-Automated Version (Excel)
available at: http://facpub.stjohns.edu/~ortizs/CLIM/CLIM-Basic.xls

Culture-Language Interpretive Matrix – Tutorial on Instruction and Interpretation in (PowerPoint)
available at: http://facpub.stjohns.edu/~ortizs/CLIM/CLIM-Instructions.ppt

Culture-Language Interpretive Matrix – General in (Word)
available at: http://facpub.stjohns.edu/~ortizs/CLIM/CLIM-General.doc

Culture-Language Test Classifications Reference List: Complete (Word)
available at: http://facpub.stjohns.edu/~ortizs/CLIM/CLTC-Reference-List.doc

Culture-Language Interpretive Matrix – Sample Validity Statements (Word)
available at: http://facpub.stjohns.edu/~ortizs/CLIM/CLIM-Interpretive-Statements.doc

Sample Report Using C-LIM – Case of Carlos – Identified as SLD (Word)
available at: http://facpub.stjohns.edu/~ortizs/CLIM/Sample-Report-Carlos-Yes-LD.doc

Sample Report Using C-LIM – Case of Maria – Not Identified as SLD (Word)
available at: http://facpub.stjohns.edu/~ortizs/CLIM/Sample-Report-Maria-No-LD.doc

The Culture-Language Interpretive Matrix (C-LIM)

Addressing test score validity for ELLs

Translation of Research into Practice

1. The use of various traditional methods for evaluating ELLs, including testing in the dominant language, modified testing, nonverbal testing, or testing in the native language do not ensure valid results and provide no mechanism for determining whether results are valid, let alone what they might mean or signify.
2. The pattern of ELL test performance, when tests are administered in English, has been established by research and is predictable and based on the examinee's degree of English language proficiency and acculturative experiences/opportunities as compared to native English speakers.
3. The use of research on ELL test performance, when tests are administered in English, provides the only current method for applying evidence to determine the extent to which obtained results are **likely valid (a minimal or only contributory influence of cultural and linguistic factors)**, **possibly valid (minimal or contributory influence of cultural and linguistic factors but which requires additional evidence from native language evaluation),** or **likely invalid (a primary influence of cultural and linguistic factors)**.
4. The principles of ELL test performance as established by research are the foundations upon which the C-LIM is based and serve as a de facto norm sample for the purposes of comparing test results of individual ELLs to the performance of a group of average ELLs with a specific focus on the attenuating influence of cultural and linguistic factors.

The Culture-Language Interpretive Matrix (C-LIM)

GENERAL RULES AND GUIDANCE FOR EVALUATION OF TEST SCORE VALIDITY

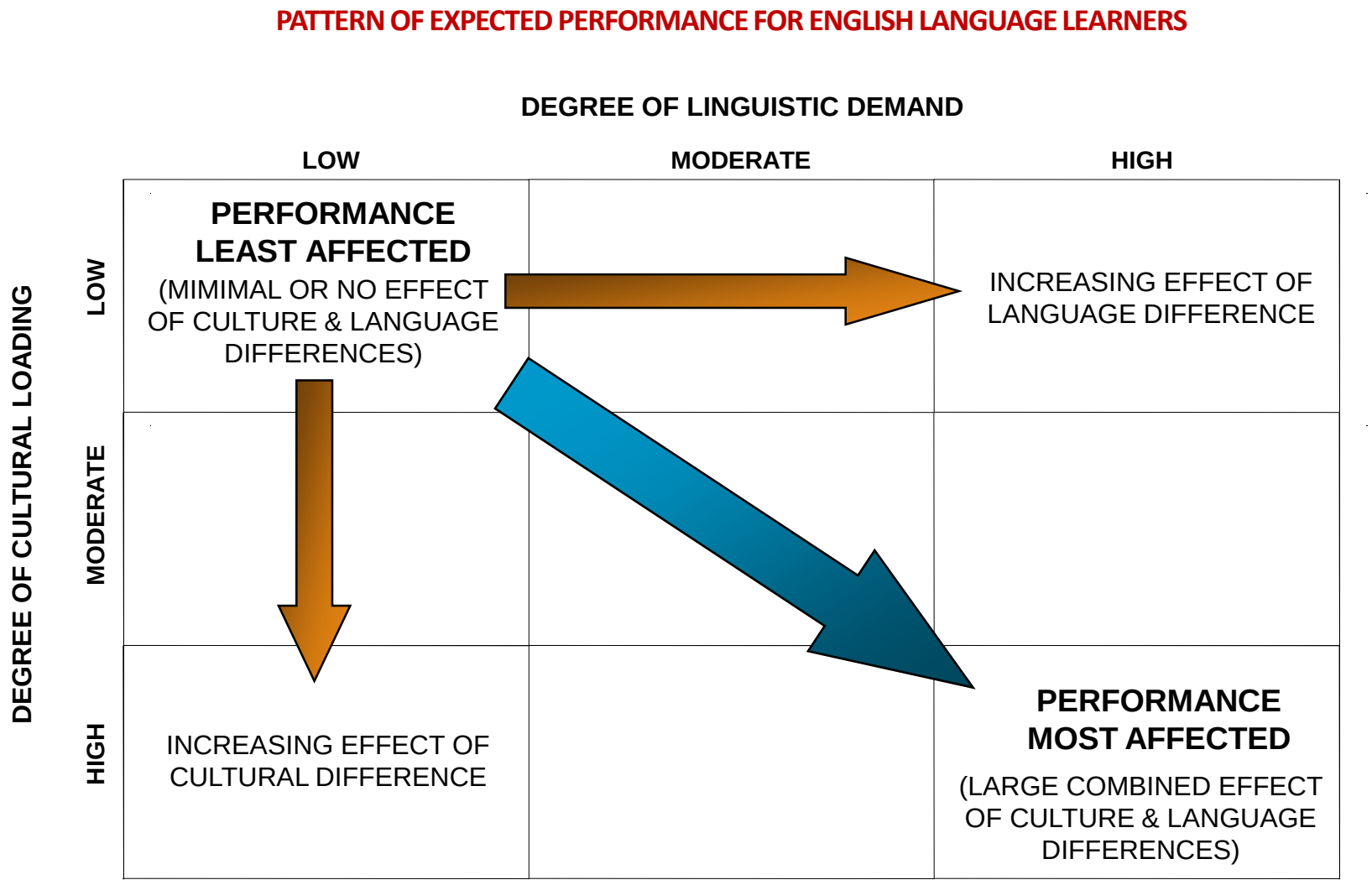
There are two basic criteria that, when both are met, provide evidence to suggest that test performance reflects the primary influence of cultural and linguistic factors and not actual ability, or lack thereof. These criteria are:

1. There exists a general, overall pattern of decline in the scores from left to right and diagonally across the matrix where performance is highest on the less linguistically demanding/culturally loaded tests (low/low cells) and performance is lowest on the more linguistically demanding/culturally loaded tests (high/high cells), and;

2. The magnitude of the aggregate test scores across the matrix for all cells fall within or above the expected range of difference (shaded area around the line) determined to be most representative of the examinee's background and development relative to the sample on whom the test was normed.

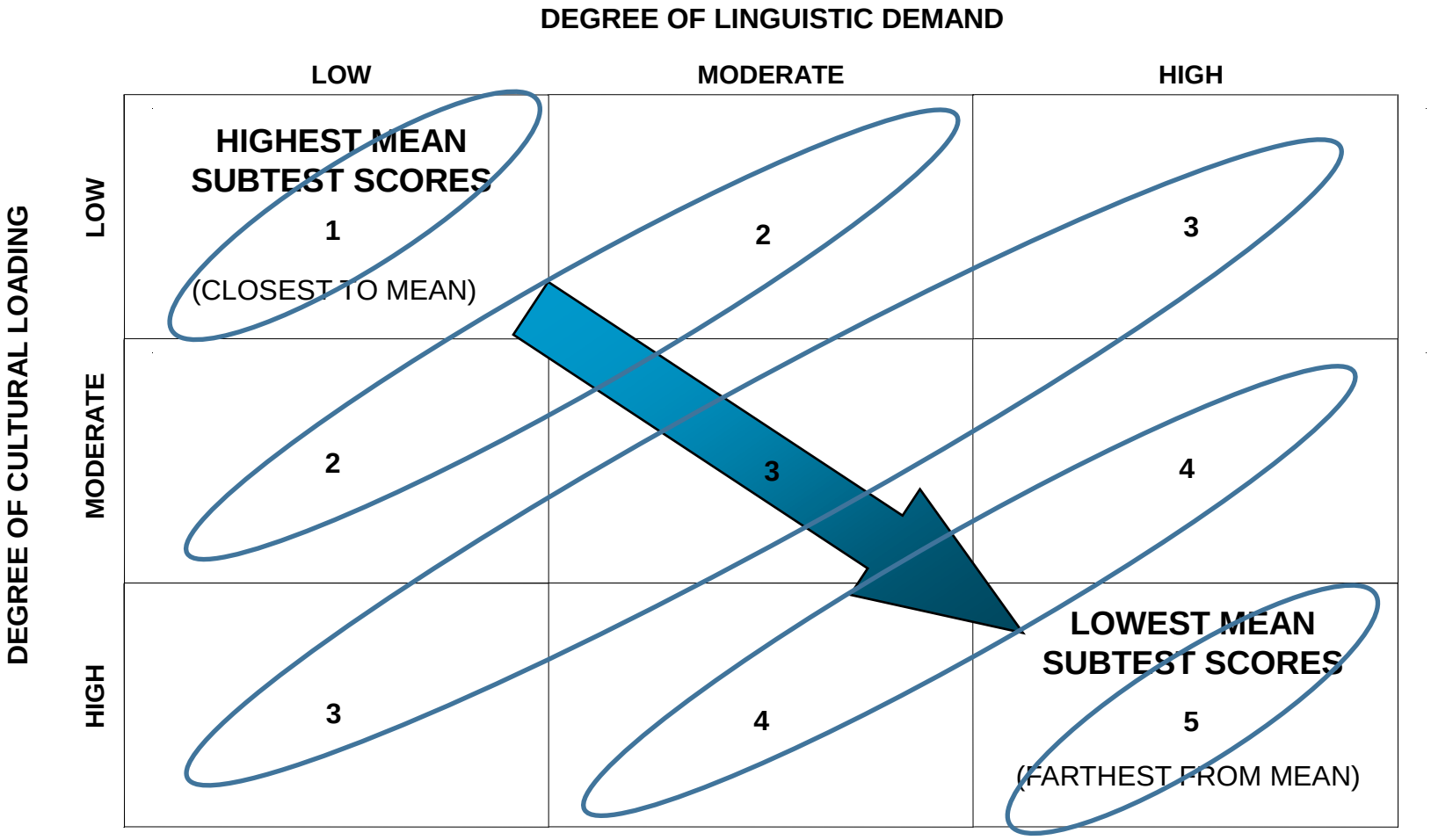
When both criteria are observed, it may be concluded that the test scores are likely to have been influenced primarily by the presence of cultural/linguistic variables and therefore are not likely to be valid and should not be interpreted.

Application of Research as Foundations for the Cultural and Linguistic Classification of Tests and Culture-Language Interpretive Matrix



Application of Research as Foundations for the Cultural and Linguistic Classification of Tests and Culture-Language Interpretive Matrix

PATTERN OF EXPECTED PERFORMANCE FOR ENGLISH LANGUAGE LEARNERS



Research Foundations for ELL Evaluation

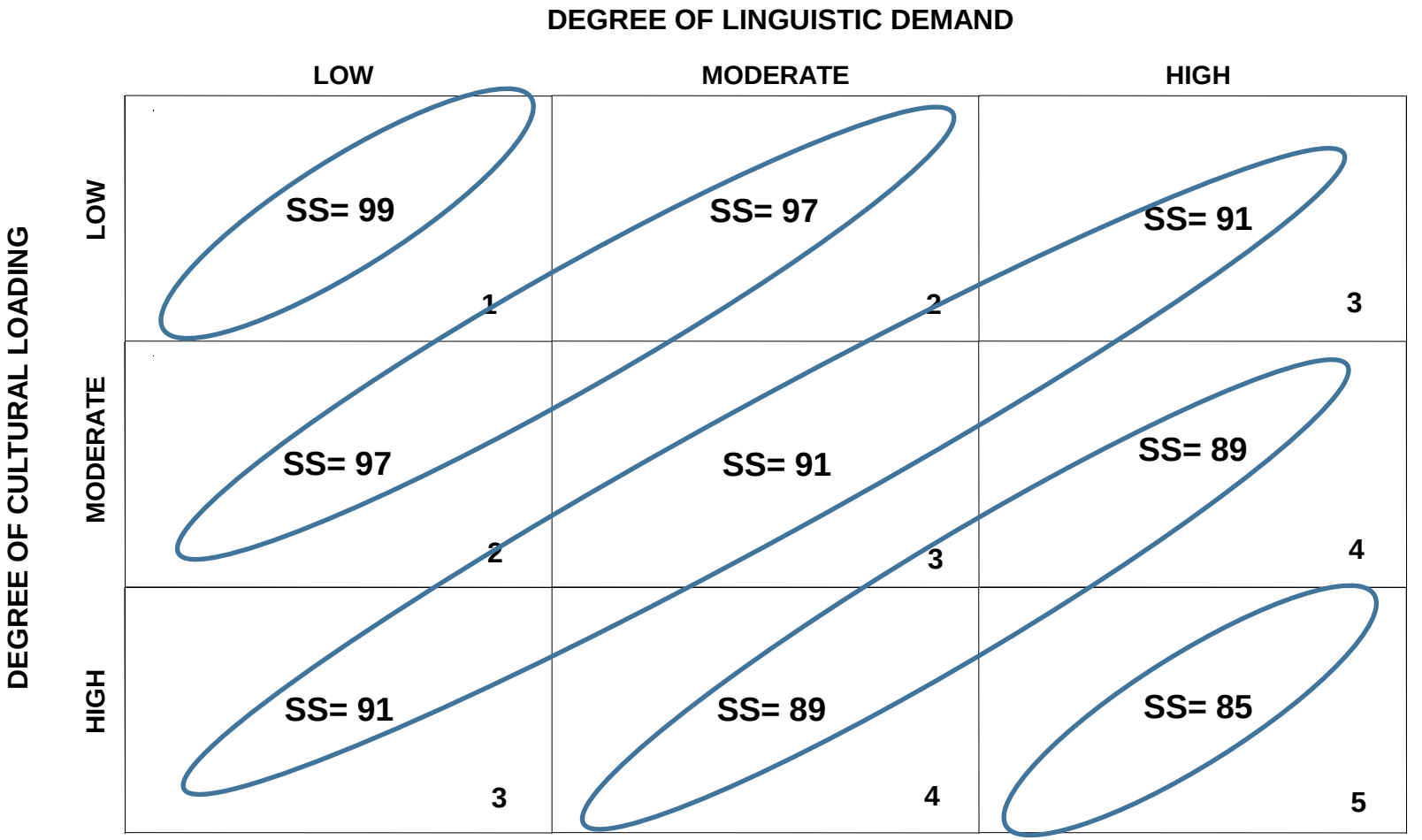
Principle 3: ELL performance is moderated by linguistic/acculturative variables

	Hispanic Group (Mercer, 1972)	Hispanic Group (Vukovich & Figueroa, 1982)	ESL Group (Cummins, 1982)	Bilingual Group (Nieves-Brull, 2006)		
Subtest Name	**Scale Score**	**Scale Score**	**Scale Score**	**Scale Score**	**Grand Mean**	**C-LIM Tier**
Information	7.5	7.8	5.1	7.2	85	5
Vocabulary	8.0	8.3	6.1	7.5	87	5
Similarities	7.6	8.8	6.4	8.2	89	4
Comprehension	7.8	9.0	6.7	8.0	89	4
Digit Span	8.3	8.5	7.3	*	90	3
Arithmetic	8.7	9.4	7.4	7.8	92	3
Picture Arrangement	9.0	10.3	8.0	9.2	96	3
Block Design	9.5	10.8	8.0	9.4	97	2
Object Assembly	9.6	10.7	8.4	9.3	98	2
Picture Completion	9.7	9.9	8.7	9.5	97	1
Coding	9.6	10.9	8.9	9.6	99	1

**Data for this subtest were not reported in the study.*

Application of Research as Foundations for the Cultural and Linguistic Classification of Tests and Culture-Language Interpretive Matrix

SAMPLE OF RESEARCH-BASED MEANS REGARDING EXPECTED PERFORMANCE FOR ENGLISH LEARNERS



Because research is conducted with highly proficient ELs, these values represent performance only for "slightly different" individuals. Those with less English proficiency will score proportionally lower.

Application of Research as Foundations for the Cultural and Linguistic Classification of Tests and Culture-Language Interpretive Matrix

RESEARCH-BASED MEANS REGARDING EXPECTED PERFORMANCE FOR ELs BY DEGREE OF DIFFERENCE

Degree of Linguistic Demand

Degree of Cultural Loading	Low	Moderate	High
Low	Slightly Different: 3-5 points Moderately Different: 5-7 points Markedly Different: 7-10 points	Slightly Different: 5-7 points Moderately Different: 7-10 points Markedly Different: 10-15 points	Slightly Different: 7-10 points Moderately Different: 10-15 points Markedly Different: 15-20 points
Moderate	Slightly Different: 5-7 points Moderately Different: 7-10 points Markedly Different: 10-15 points	Slightly Different: 7-10 points Moderately Different: 10-15 points Markedly Different: 15-20 points	Slightly Different: 10-15 points Moderately Different: 15-20 points Markedly Different: 20-25 points
High	Slightly Different: 7-10 points Moderately Different: 10-15 points Markedly Different: 15-20 points	Slightly Different: 10-15 points Moderately Different: 15-20 points Markedly Different: 20-25 points	Slightly Different: 15-20 points Moderately Different: 20-25 points Markedly Different: 25-35 points

***Slightly Different**: Includes individuals with very high levels of English language proficiency (e.g., CALP) and high acculturation, but still not entirely comparable to mainstream U.S. English speakers. Examples include individuals who are third generation in the U.S., have well educated/higher SES parents, have attended dual-language program for at least 6-7 years, or demonstrate native or near native-like proficiency in English language conversation and solid literacy skills. (Not a common category)*

***Moderately Different**: Includes individuals with moderate to higher levels of English language proficiency (e.g., advanced BICS/emerging CALP) and typical EL acculturative learning experiences. Examples include individuals who were born or came early to the U.S. with limited English speaking parents, usually from low to very low SES with parent's having low or limited literacy even in their own language, generally received formal education in English only or primarily in English since starting school.*

***Markedly Different**: Includes individuals with low to very low levels of English language proficiency (e.g., early BICS) or very limited acculturative learning experiences due to unusual influences on development. Examples include extremely low and limited parental SES and education, recently arrival in the U.S. or residence for in the U.S. 3 years or less, lack of prior formal education, exposure to trauma, violence, abuse, neglect, time spent in refugee or resettlement camps, changes in or multiple early languages.*

Research Foundations of the C-LIM
Additional Issues in Evaluation of Test Score Patterns

Evaluation of test score validity, particularly in cases where results are "possibly valid," includes considerations such as:

1. Is the Tiered graph consistent with the main Culture-Language graph or the other secondary (language-only/culture-only) graphs?

2. Is there any variability in the scores that form the aggregate in a particular cell that may be masking low performance?

3. Is the pattern of scores consistent with a developmental explanation of the examinee's educational program and experiences?

4. Is the pattern of scores consistent with a developmental explanation of the examinee's linguistic/acculturative learning experiences?

Evaluation of results using all graphs, including secondary ones, identification of score variability in relation to CHC domains or task characteristics, and evaluation of educational, cultural, and linguistic developmental experiences assists in determining the most likely cause of score patterns and overall test score validity.

The Culture-Language Interpretive Matrix (C-LIM)

RANGE OF POSSIBLE OUTCOMES WHEN EVALUATING TEST SCORES WITHIN C-LIM

Condition A: Overall pattern generally appears to decline across all cells and all cell aggregate scores within or above shaded range—test scores **likely invalid**, cultural/linguistic factors are primary influences, but examinee likely has average/higher ability as data do not support deficits, and further evaluation via testing is unnecessary.

Condition B: Overall pattern generally appears to decline across all cells but at least one cell aggregate (or more) is below shaded range—test scores **possibly valid**, cultural/linguistic factors are contributory influences, and further evaluation, including in the native language, is necessary to establish true weaknesses in a given domain.

Condition C: Overall pattern does not appear to decline across all cells and all cell aggregate scores within or above average range—test scores **likely valid**, cultural/linguistic factors are minimal influences, and further evaluation may be unnecessary if no weaknesses exist in any domain.

Condition D: Overall pattern does not appear to decline across all cells and at least one cell aggregate (or more) is below average range—test scores **possibly valid**, cultural/linguistic factors are minimal influences, and further evaluation, including in the native language, is necessary to establish true weaknesses in a given domain.

The Culture-Language Interpretive Matrix (C-LIM)

RANGE OF POSSIBLE OUTCOMES WHEN EVALUATING TEST SCORES WITHIN C-LIM

	A general, overall pattern of decline exists?	All scores within or above the expected range?	All scores within or above the average range?	Degree of influence of cultural and linguistic factors	Likelihood that test scores are valid indicators of ability?
Condition A	Yes	Yes	No	Primary	Unlikely
Condition B	Yes	No	No	Contributory	Possibly*
Condition C	No	Yes	Yes	Minimal	Likely
Condition D	No	No	No	Minimal	Possibly*

**Determination regarding the validity of test scores that are below the expected and average ranges requires additional data and information, particularly results from native language evaluation, qualitative evaluation and analysis, and data from a strong pre-referral process (e.g., progress monitoring data).*

C-LIM Guidelines for Evaluating Test Scores

CONDITION A: General declining pattern, all scores within or above expected range.

Culture-Language Interpretive Matrix - Analyzer and Data Entry

Name: *Sydney - DAS-II* Age: *9 years 8 month(s)* Grade: *4* Date: *1/1/2015*

DEGREE OF LINGUISTIC DEMAND

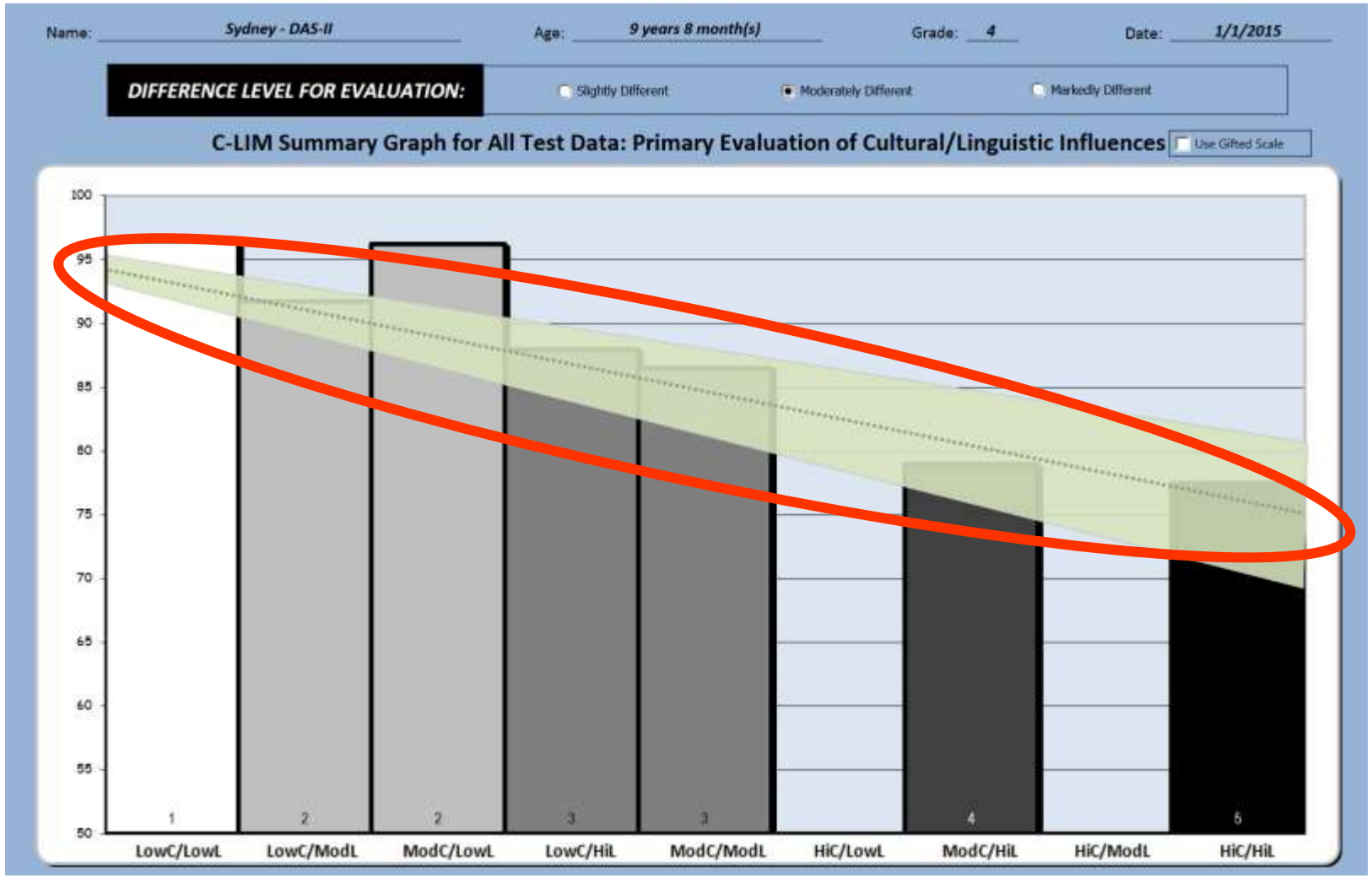
DEGREE OF CULTURAL LOADING

	LOW		MODERATE		HIGH	
		Score		Score		Score
LOW	DAS-II Copying		DAS-II Recall of Digits-Backward	49 99	DAS-II Recall of Digits-Forward	42 88
	DAS-II Matching Letter-Like Forms		DAS-II Speed of Information Processing	40 85		
	DAS-II Matrices	45 93				
	DAS-II Pattern Construction	54 106				
	DAS-II Recall of Designs					
	DAS-II Sequential & Quantitative Reasoning	44 91				
	Cell Average =	97	Cell Average =	92	Cell Average =	88
		Score		Score		Score
MODERATE	DAS-II Picture Similarities		DAS-II Early Number Concepts		DAS-II Phonological Processing	36 79
	DAS-II Recall of Objects-Delayed		DAS-II Rapid Naming	41 87		
	DAS-II Recall of Objects-Immediate	45 93	DAS-II Recognition of Pictures			
	DAS-II Recognition of Pictures	50 100				
	Cell Average =	96	Cell Average =	87	Cell Average =	79
		Score		Score		Score
HIGH			DAS-II Naming Vocabulary		DAS-II Word Similarities	31 72
			DAS-II Verbal Comprehension		DAS-II [illegible]	39 84
	Cell Average =		Cell Average =		Cell Average =	78

CULTURE/LANGUAGE INFLUENCE: PRIMARY – all test scores are UNLIKELY to be valid.

C-LIM Guidelines for Evaluating Test Scores

CONDITION A: General declining pattern, all scores within or above expected range.



CULTURE/LANGUAGE INFLUENCE: PRIMARY – all test scores are UNLIKELY to be valid.

C-LIM Guidelines for Evaluating Test Scores

CONDITION B: Generally declining pattern, one or more scores below expected range.

Culture-Language Interpretive Matrix - Analyzer and Data Entry

Name: *Carmen - KABC-II* Age: *9 years 9 month(s)* Grade: *4* Date: *1/1/2015*

DEGREE OF LINGUISTIC DEMAND

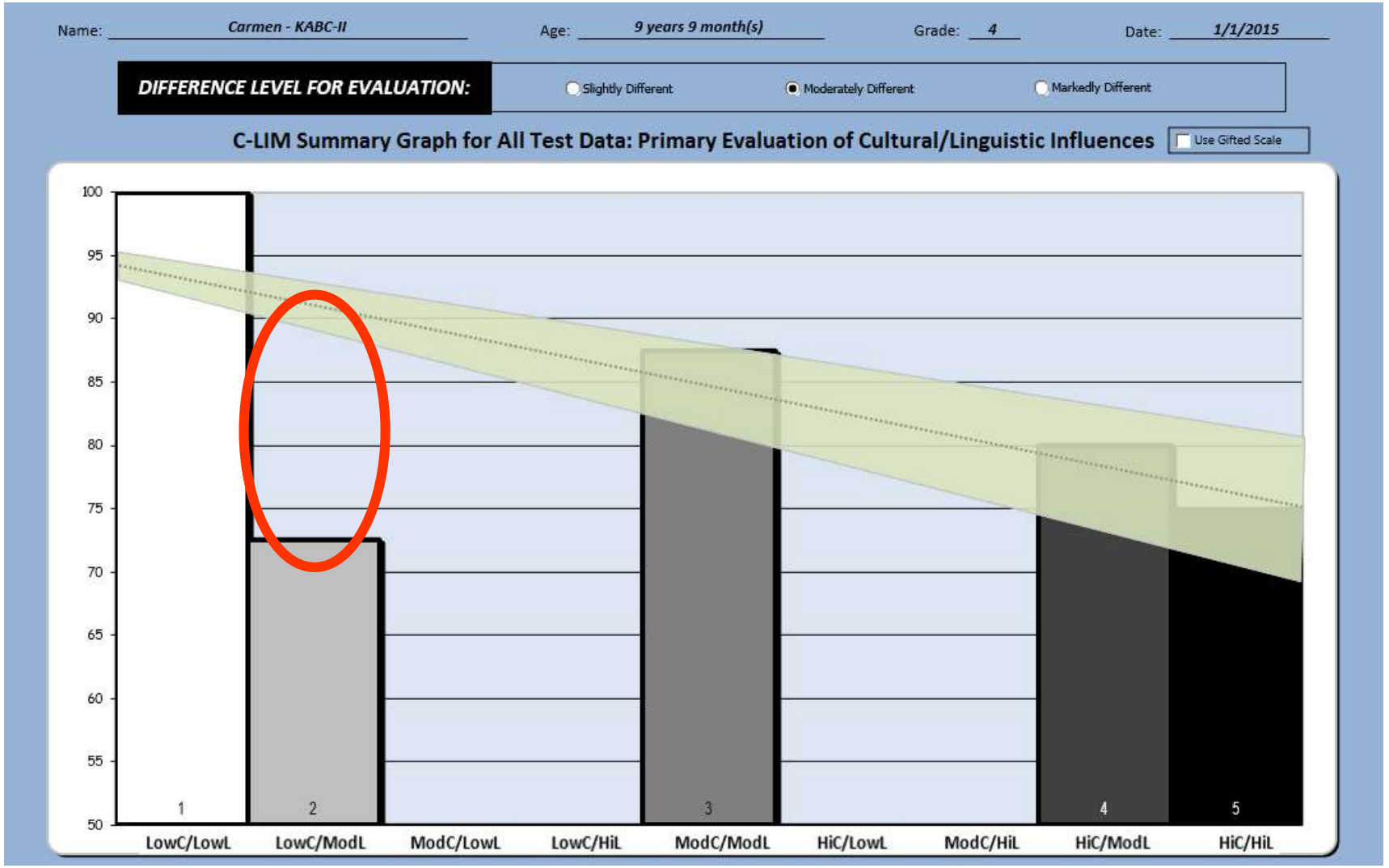
DEGREE OF CULTURAL LOADING

	LOW		MODERATE		HIGH	
		Score		Score		Score
LOW	KABC-II Atlantis	10 *100*	KABC-II Block Counting			
	KABC-II Atlantis Delayed		KABC-II Number Recall	4 *70*		
	KABC-II Face Recognition		KABC-II Rebus	5 *75*		
	KABC-II Hand Movements		KABC-II Rebus Delayed			
	KABC-II Pattern Reasoning (7-18 years)	11 *105*				
	KABC-II Triangles	9 *95*				
	Cell Average =	**100**	Cell Average =	**73**	Cell Average =	
MODERATE		Score		Score		Score
			KABC-II Conceptual Thinking			
			KABC-II Rover	7 *85*		
			KABC-II Word Order	8 *90*		
	Cell Average =		Cell Average =	**88**	Cell Average =	
HIGH	KABC-II Gestalt Closure	Score	KABC-II Story Completion (7-18 years)	Score 6 *80*	KABC-II Expressive Vocabulary	Score
					KABC-II Riddles	5 *75*
					KABC-II Verbal Knowledge	5 *75*
	Cell Average =		Cell Average =	**80**	Cell Average =	**75**

CULTURE/LANGUAGE INFLUENCE: CONTRIBUTORY – low test scores are POSSIBLY valid.

C-LIM Guidelines for Evaluating Test Scores

CONDITION B: Generally declining pattern, one or more scores below expected range.



CULTURE/LANGUAGE INFLUENCE: **CONTRIBUTORY – low test scores are POSSIBLY valid.**

C-LIM Guidelines for Evaluating Test Scores

CONDITION C: No declining pattern, all scores within or above average range.

Culture-Language Interpretive Matrix - Analyzer and Data Entry

Name: *Benjamin - WISC-V* Age: *10 years 8 month(s)* Grade: *5* Date: *1/1/2015*

DEGREE OF LINGUISTIC DEMAND

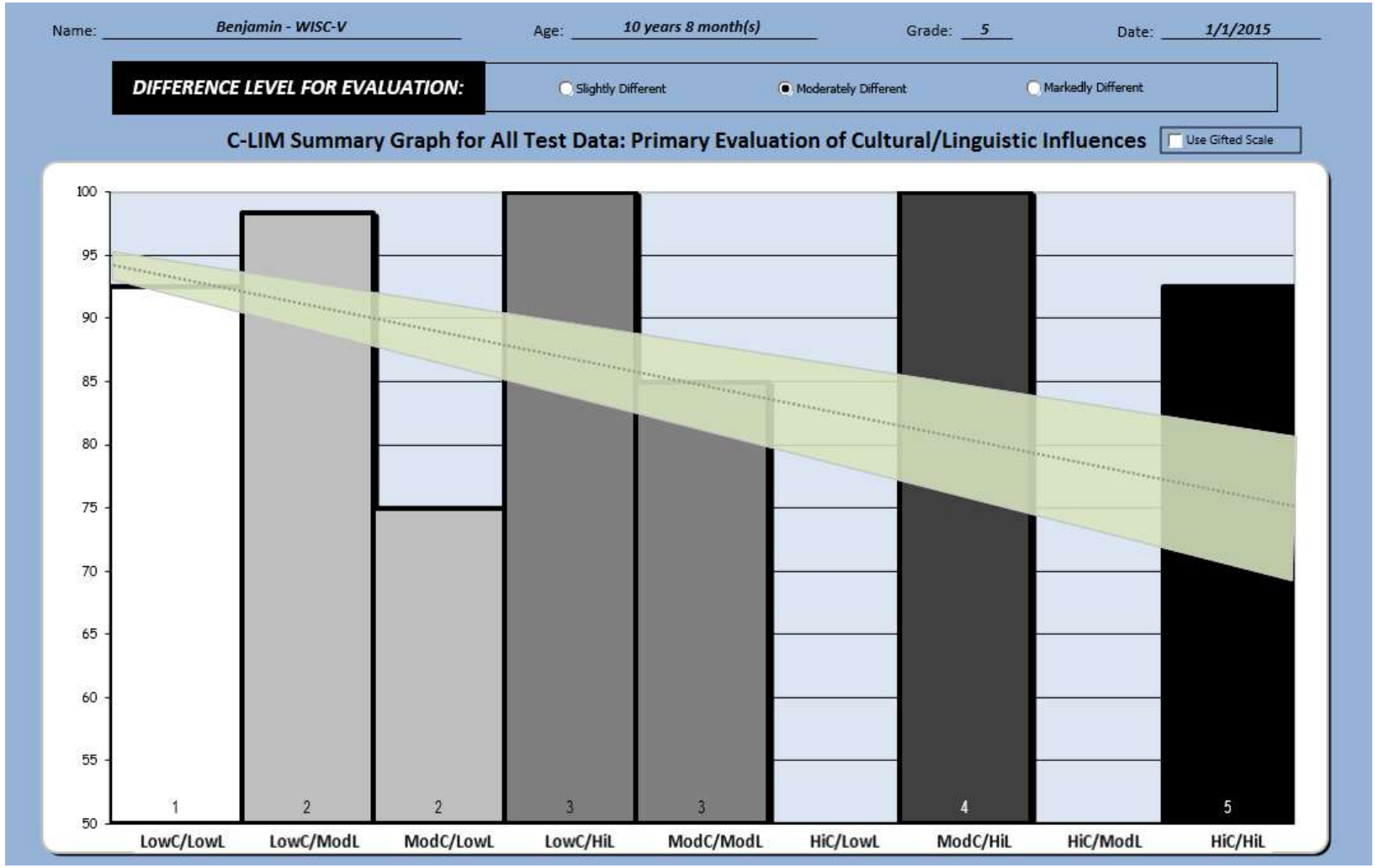
DEGREE OF CULTURAL LOADING

	LOW		MODERATE		HIGH	
		Score		Score		Score
LOW	WISC-V Cancellation	9 95	WISC-V Block Design	10 100	WISC-V Digit Span	
	WISC-V Matrix Reasoning	8 90	WISC-V Coding	9 95	WISC-V Letter-Number Sequencing	10 100
	WISC-V Naming Speed Quantity		WISC-V Delayed Symbol Translation			
	WISC-V Visual Puzzles		WISC-V Immediate Symbol Translation			
			WISC-V Picture Span			
			WISC-V Recognition Symbol Translation			
			WISC-V Symbol Search	10 100		
	Cell Average =	93	Cell Average =	98	Cell Average =	100
		Score		Score		Score
MODERATE	WISC-V Picture Concepts	5 75	WISC-V Arithmetic	7 85	WISC-V Comprehension	10 100
			WISC-V Figure Weights			
			WISC-V Naming Speed Literacy			
	Cell Average =	75	Cell Average =	85	Cell Average =	100
		Score		Score		Score
HIGH					WISC-V Information	
					WISC-V Similarities	9 95
					WISC-V Vocabulary	8 90
	Cell Average =		Cell Average =		Cell Average =	93

CULTURE/LANGUAGE INFLUENCE: MINIMAL – all test scores are LIKELY to be valid.

C-LIM Guidelines for Evaluating Test Scores

CONDITION C: No declining pattern, all scores within or above average range.



CULTURE/LANGUAGE INFLUENCE: MINIMAL – all test scores are LIKELY to be valid.

C-LIM Guidelines for Evaluating Test Scores

CONDITION D: No declining pattern, one or more scores below average range.

Culture-Language Interpretive Matrix - Analyzer and Data Entry

Name: *Katrina - WJ IV* Age: *8 years 6 month(s)* Grade: *3* Date: *1/1/2015*

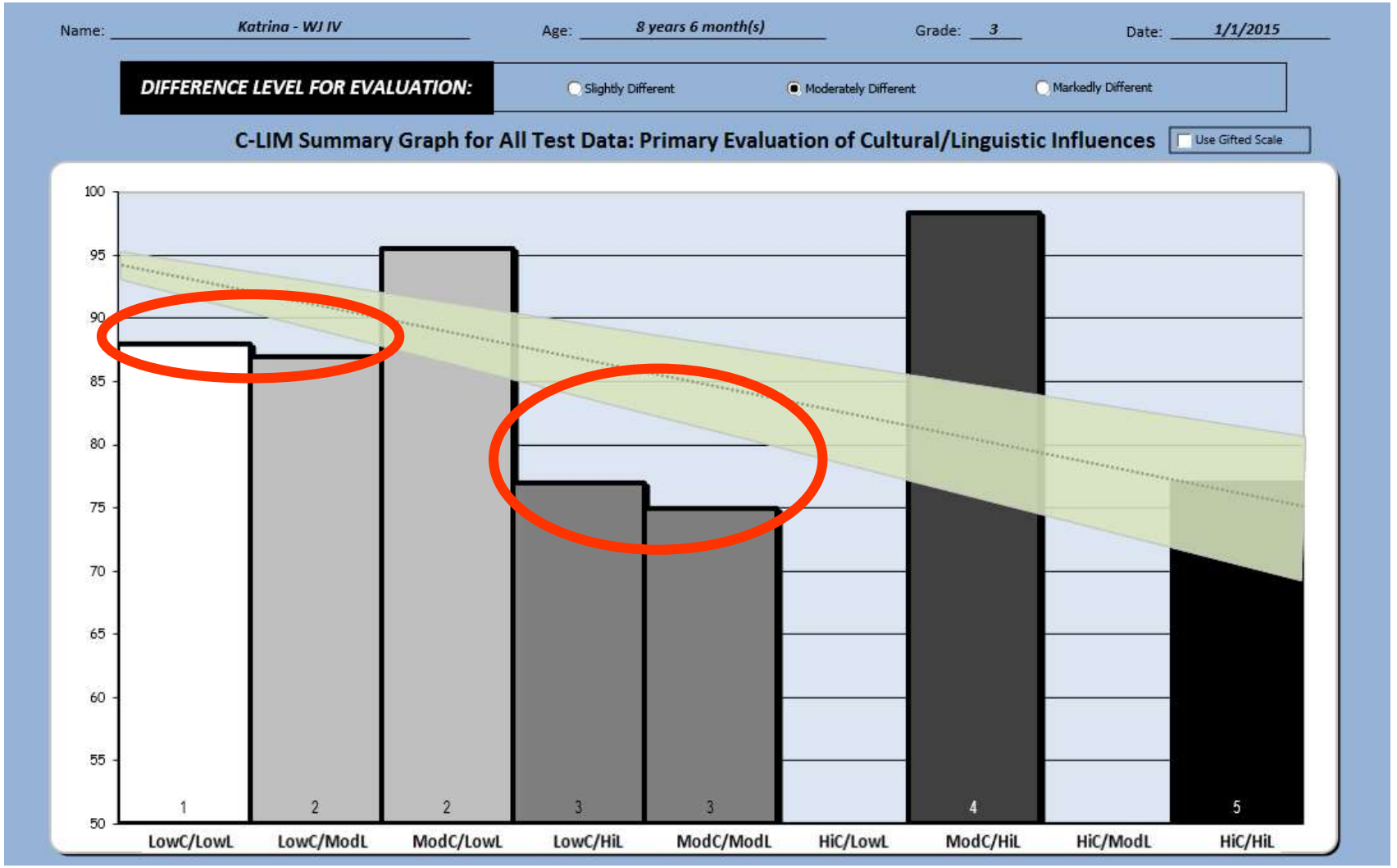
DEGREE OF LINGUISTIC DEMAND

DEGREE OF CULTURAL LOADING	LOW		MODERATE		HIGH	
		Score		Score		Score
LOW	WJ IV COG Number Series	82 82	WJ IV COG Analysis-Synthesis	87 87	WJ IV COG Concept Formation	77 77
	WJ IV COG Number-Pattern Matching		WJ IV COG Numbers Reversed	87 87	WJ IV COG Object-Number Sequencing	
	WJ IV COG Pair Cancellation	110 110				
	WJ IV COG Visualization	72 72				
	Cell Average =	88	Cell Average =	87	Cell Average =	77
		Score		Score		Score
MODERATE	WJ IV COG Letter-Pattern Matching	99 99	WJ IV COG Nonword Repetition	80 80	WJ IV COG Memory for Words	104 104
	WJ IV COG Picture Recognition	92 92	WJ IV COG Visual Auditory Learning	70 70	WJ IV COG Phonological Processing	99 99
					WJ IV COG Verbal Attention	92 92
					WJ IV OL Sentence Repetition	
	Cell Average =	96	Cell Average =	75	Cell Average =	98
		Score		Score		Score
HIGH			WJ IV OL Picture Vocabulary		WJ IV COG General Information	80 80
					WJ IV COG Oral Vocabulary	78 78
					WJ IV COG Story Recall	73 73
	Cell Average =		Cell Average =		Cell Average =	77

CULTURE/LANGUAGE INFLUENCE: MINIMAL – low test scores are POSSIBLY valid.

C-LIM Guidelines for Evaluating Test Scores

CONDITION D: No declining pattern, one or more scores below average range.



CULTURE/LANGUAGE INFLUENCE: MINIMAL – low test scores are POSSIBLY valid.

C-LIM Additional Interpretive Issues

KABC-II DATA FOR TRAN (ENGLISH)

Culture-Language Interpretive Matrix - Analyzer and Data Entry

Name: *Tran - KABC-II* Age: *10 years 10 month(s)* Grade: *4* Date: *1/1/2015*

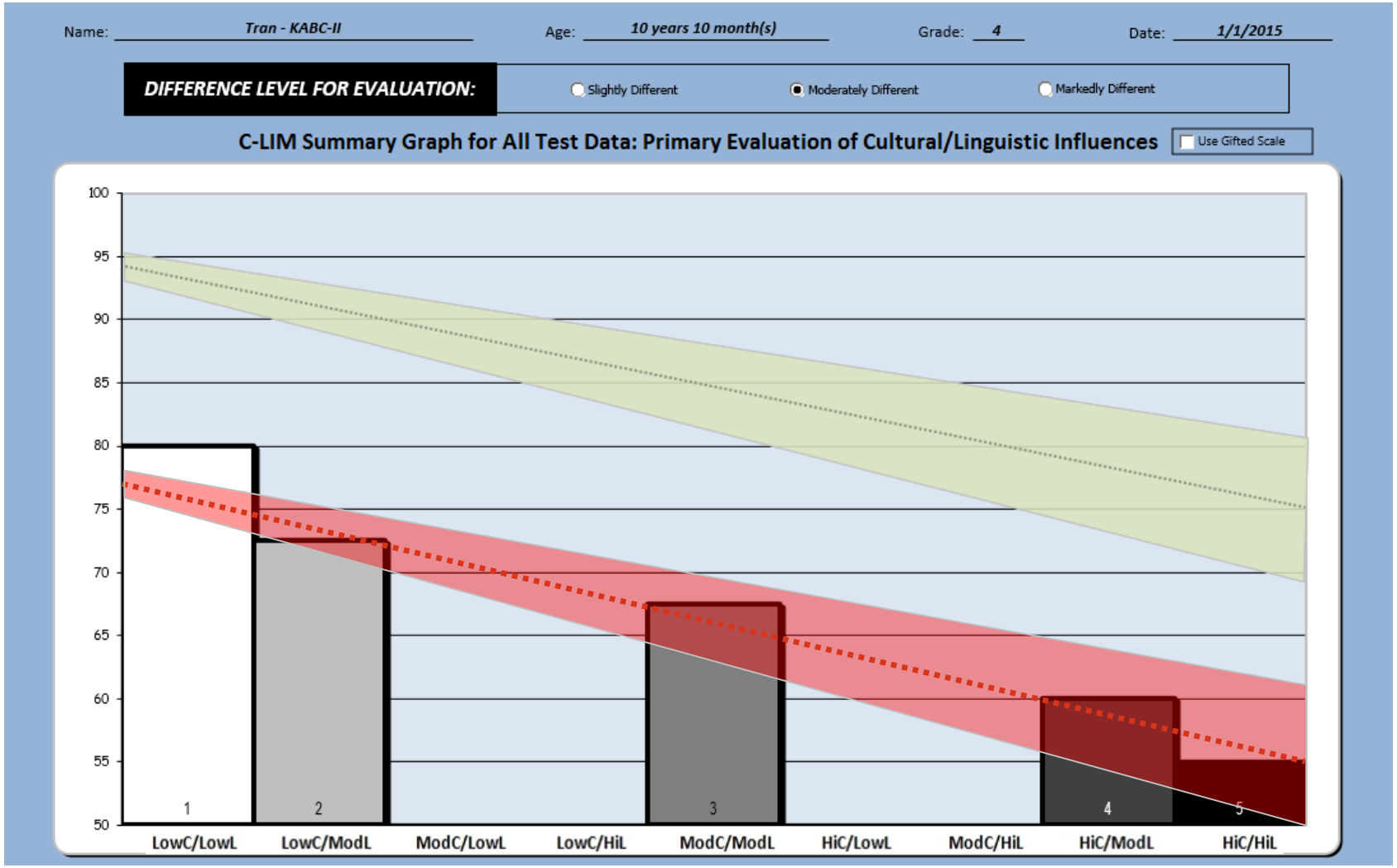
DEGREE OF LINGUISTIC DEMAND

DEGREE OF CULTURAL LOADING

	LOW		MODERATE		HIGH	
		Score		Score		Score
LOW	KABC-II Atlantis	6 80	KABC-II Block Counting			
	KABC-II Atlantis Delayed		KABC-II Number Recall	5 75		
	KABC-II Face Recognition		KABC-II Rebus	4 70		
	KABC-II Hand Movements		KABC-II Rebus Delayed			
	KABC-II Pattern Reasoning (7-18 years)	5 75				
	KABC-II Triangles	7 85				
	Cell Average =	80	Cell Average =	73	Cell Average =	
MODERATE		Score		Score		Score
			KABC-II Conceptual Thinking			
			KABC-II Rover	3 65		
			KABC-II Word C	4 70		
	Cell Average =		Cell Average =	68	Cell Average =	
HIGH		Score		Score		Score
	KABC-II Gestalt Closure		KABC-II Story Completion (7-18 years)	2 60	KABC-II Expressive Vocabulary	
					KABC-II Riddl	1 55
					KABC-II Ver	1 55
	Cell Average =		Cell Average =	60	Cell Average =	55

C-LIM Additional Interpretive Issues

KABC-II DATA FOR TRAN (ENGLISH)



CONDITION B: Generally declining pattern, one or more scores below expected range.

CULTURE/LANGUAGE INFLUENCE: CONTRIBUTORY – low test scores are POSSIBLY valid.

C-LIM Additional Interpretive Issues

WJ IV COG DATA FOR HADJI (ENGLISH)

Culture-Language Interpretive Matrix - Analyzer and Data Entry

Name: *Hadji - WJ IV* Age: *11 years 6 month(s)* Grade: *6* Date: *1/1/2015*

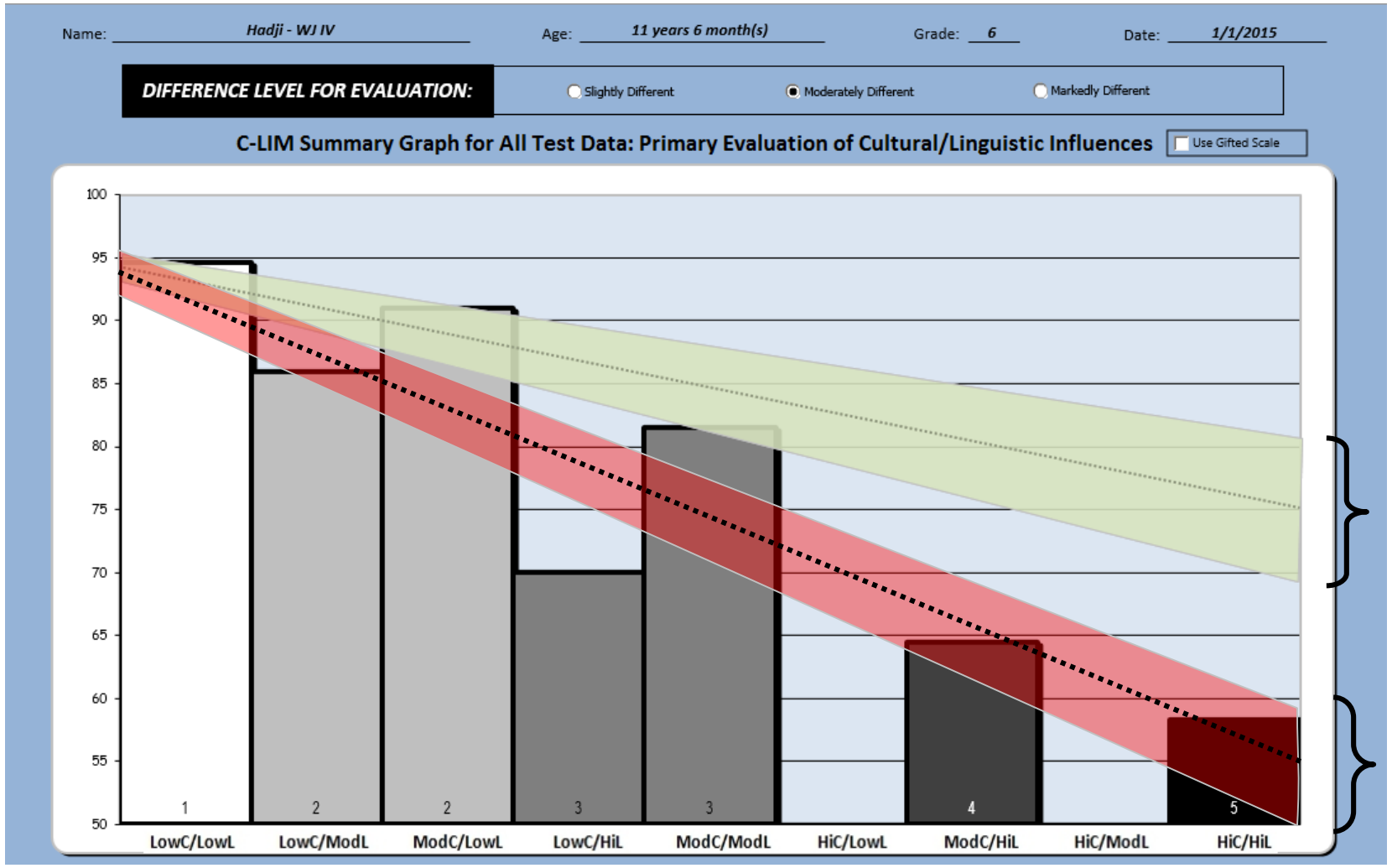
DEGREE OF LINGUISTIC DEMAND

DEGREE OF CULTURAL LOADING

	LOW		MODERATE		HIGH	
		Score		Score		Score
LOW	WJ IV COG Number Series	99 99	WJ IV COG Numbers Reversed	86 86	WJ IV COG Concept Formation	70 70
	WJ IV COG Pair Cancellation	94 94				
	WJ IV COG Visualization	91 91				
	Cell Average =	95	Cell Average =	86	Cell Average =	70
		Score		Score		Score
MODERATE	WJ IV COG Letter-Pattern Matching	89 89	WJ IV COG Nonword Repetition	86 86	WJ IV COG Phonological Processing	68 68
	WJ IV COG Picture Recognition	93 93	WJ IV COG Visu ditory Learning	77 77	WJ IV COG Verbal Attention	61 61
	Cell Average =	91	Cell Average =	82	Cell Average =	65
		Score		Score		Score
HIGH					WJ IV COG General Information	62 62
					WJ IV COG O cabulary	59 59
					WJ IV COG	54 54
	Cell Average =		Cell Average =		Cell Average =	58

C-LIM Additional Interpretive Issues

WJ IV COG DATA FOR HADJI (ENGLISH)



Expected rate of decline

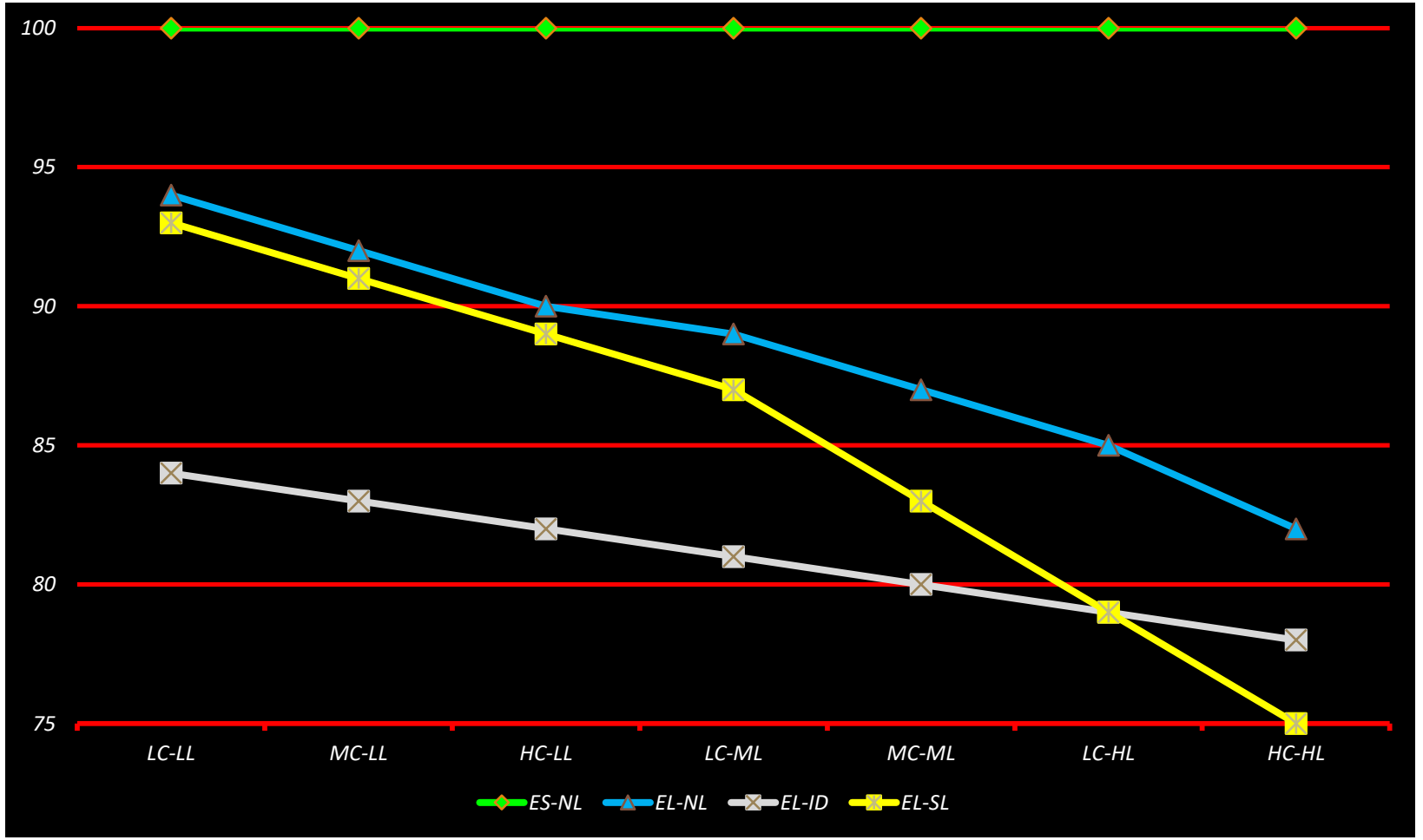
Steeper rate of decline

CONDITION B: Generally declining pattern, one or more scores below expected range.

CULTURE/LANGUAGE INFLUENCE: CONTRIBUTORY – low test scores are POSSIBLY valid.

Comparison of Patterns of Performance Among English-Speakers and English-Learners with SLD, SLI, and ID

Mean C-LIM cell aggregates for WPPSI-III subtests arranged by degree of cultural loading and linguistic demand



Source: Tychanska, J., Ortiz, S. O., Flanagan, D.P., & Terjesen, M. (2009), unpublished data..

Translating Research into Practice

Evaluation Issues and Methods	Norm sample representative of bilingual development	Measures a wider range of school-related abilities	Does not require the evaluator to be bilingual	Adheres to the test's standardized protocol	Substantial research base on bilingual performance	Sufficient to identify or diagnosis disability	Accounts for variation in bilingual development	Most likely to yield reliable and valid data and information	Provides extensive data regarding development
Modified or Altered Assessment	✗	✓	✓	✗	✗	✗	✗	✗	✗
Reduced-language Assessment	✗	✗	✓	✓	✗	✗	✗	✗	✗
Dominant Monolingual Assessment in L1: native only	✗	✓	✗	✓	✗	✗	✗	✗	✗
Dominant Monolingual Assessment in L2: English only	✗	✓	✓	✓	✓	✗	✗	✗	✗
Multilingual Assessment in L1 + L2	✗	✓	✓	✓	✓	✓	✓	✓	✓

Multilingual Assessment combined with the C-LIM resolves all validity issues, and by applying research on ELL test performance, they can be used to define and establish a "true peer" reference group for disability-based evaluations.

Practical Considerations for Addressing Validity in Disability Evaluation Procedures with ELLs

Although the C-LIM provides a “true peer” comparison base, its use must be integrated within a “best practices” framework for assessment of English learners.

Such a framework is not achieved by any of the previous methods alone (i.e., testing the limits, nonverbal testing, testing in L1), and requires consideration of other factors that are often overlooked or ignored (e.g., alterations to standard protocol renders scores invalid, only a small fraction of evaluators are bilingual, suitable tests are often not available in languages other than Spanish, ELL students in the U.S. are “bilingual” not monolingual, etc.). Moreover, research on ELL test performance exists only for tests administered in English, not in the native language.

Therefore, in addition to the need to be systematic and research-based, a best practice framework for assessment of English learners must take into account these and other practical issues so as to be applicable and useful for all evaluators, not merely those with bilingual skills and training.

Practical Considerations for Addressing Validity in Disability Evaluation Procedures with ELLs

1. *The usual purpose of testing is to identify deficits in ability (i.e., low scores)*
2. *Validity is more of a concern for low scores than average/higher scores because:*
    - *Test performances in the average range are NOT likely a chance finding and strongly suggests average ability (i.e., no deficits in ability)*
    - *Test performances that are below average MAY be a chance finding because of experiential or developmental differences and thus do not automatically confirm below average ability (i.e., possible deficits in ability)*
3. *Therefore, testing in one language only (English or native language) means that:*
    - *It can be determined that a student DOES NOT have a disability (i.e., if all scores are average or higher, they are very likely to be valid)*
    - *It CANNOT be determined if the student has a disability (i.e., low scores must be validated as true indicators of deficit ability)*
4. *Testing in both languages (English and native language) is necessary to determine disability*
    - *Testing requires confirmation that deficits are not language-specific and exist in both languages (although low performance in both can result from other factors)*
5. *All low test scores, whether in English or the native language, must be validated*
    - *Low scores from testing in English can be validated via research underlying the C-LIM*
    - *Low scores from testing in the native language cannot be validated with research*

Practical Considerations for Addressing Validity in Disability Evaluation Procedures with ELLs

Given the preceding considerations, the most practical and defensible general approach in evaluating ELLs would be:

- *Test in English first and if all test scores indicate strengths (average or higher) a disability is not likely and thus no further testing is necessary*
- *If some scores from testing in English indicate weaknesses, re-test those areas in the native language to support them as areas of true weakness*
- *Use all other case data and information to serve as the context by which to evaluate the test scores and ensure ecological validity to conclusions*

When combined with the C-LIM, this approach provides an efficient, research-based, and IDEA-compliant process that makes best use of available resources for evaluation consistent with current standards as it permits ANY evaluator to begin (and in some cases, complete) the testing without being bilingual or requiring outside assistance.

A Best Practice Framework for Evaluation and Disability Testing with ELLs

Step 1. Evaluate construct validity in all areas in English (exclusion of cultural/linguistic factors)

- *Test in English first and use C-LIM to evaluate scores. If all scores indicate strengths (average or higher) a disability is not likely and no further testing is necessary. If any scores suggest weaknesses, continue evaluation.*

Step 2. Re-evaluate construct validity in areas of weakness in native language (cross-linguistic evidence)

- *If some scores from testing in English indicate weaknesses, re-test those areas in the native language to support them as areas of true weakness*

Step 3. Cross-validate L1 and L2 test scores with contextual factors and data (ecological validity for disability)

- *Use all other case data and information to serve as the context by which to evaluate the test scores and ensure ecological validity to conclusions*

A Guided Case Study Example of Evaluation of an English Learner for Specific Learning Disability

Evaluation of Jose Maria
Tests Used: WISC-V, WIAT-III, and WJ IV
DOE: 6/22/2016
DOB: 10/4/2006
Grade: 4

SLD Identification with an English Learner: A Case Study

Step 1. Evaluate construct validity in all areas in English (exclusion of cultural/linguistic factors)

- *Test in English first and use C-LIM to evaluate scores. If all scores indicate strengths (average or higher) a disability is not likely and no further testing is necessary. If any scores suggest weaknesses, continue evaluation.*

Step 2. Re-evaluate construct validity in areas of weakness in native language (cross-linguistic evidence)

- *If some scores from testing in English indicate weaknesses, re-test those areas in the native language to support them as areas of true weakness*

Step 3. Cross-validate L1 and L2 test scores with contextual factors and data (ecological validity for disability)

- *Use all other case data and information to serve as the context by which to evaluate the test scores and ensure ecological validity to conclusions*

SLD Identification with an English Learner: A Case Study

WISC-V/WJ IV/WIAT-III XBA DATA FOR Jose Maria
DOE: 6/22/2016 DOB: 10/4/2006 Grade: 4

WECHSLER INTELLIGENCE SCALE FOR CHILDREN-V

Verbal Comprehension Index	*76*
Similarities	5
Vocabulary	6

Fluid Reasoning Index	*82*
Matrix Reasoning	7
Figure Weights	7

Visual-Spatial Index	*95*
Block Design	9
Visual Puzzles	9

Working Memory Index	*79*
Digit Span	5
Picture Span	7

Processing Speed Index	*94*
Coding	9
Symbol Search	8

WECHSLER INDIVIDUAL ACHIEVEMENT TEST-III

Basic Reading	*94*
Word Reading	92
Pseudoword Decoding	98

Reading Comprehension	*76*
Reading Comprehension	76
Oral Reading Fluency	80

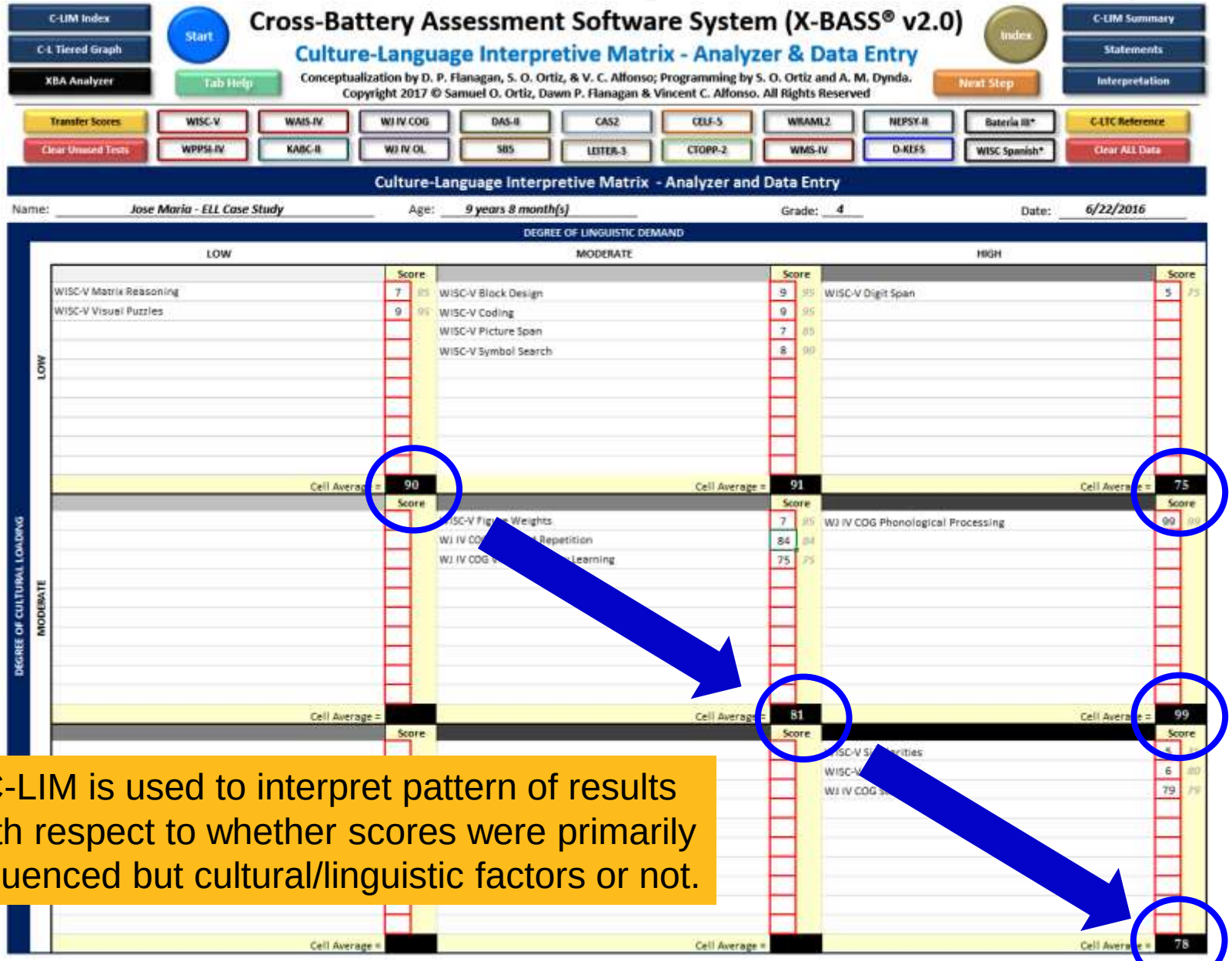
Written Expression	*92*
Spelling	100
Sentence Composition	86
Essay Composition	93

WOODCOCK JOHNSON-IV TESTS OF COGNITIVE ABILITY

Auditory Processing	*91*
Phonological Processing	99
Nonword Repetition	84

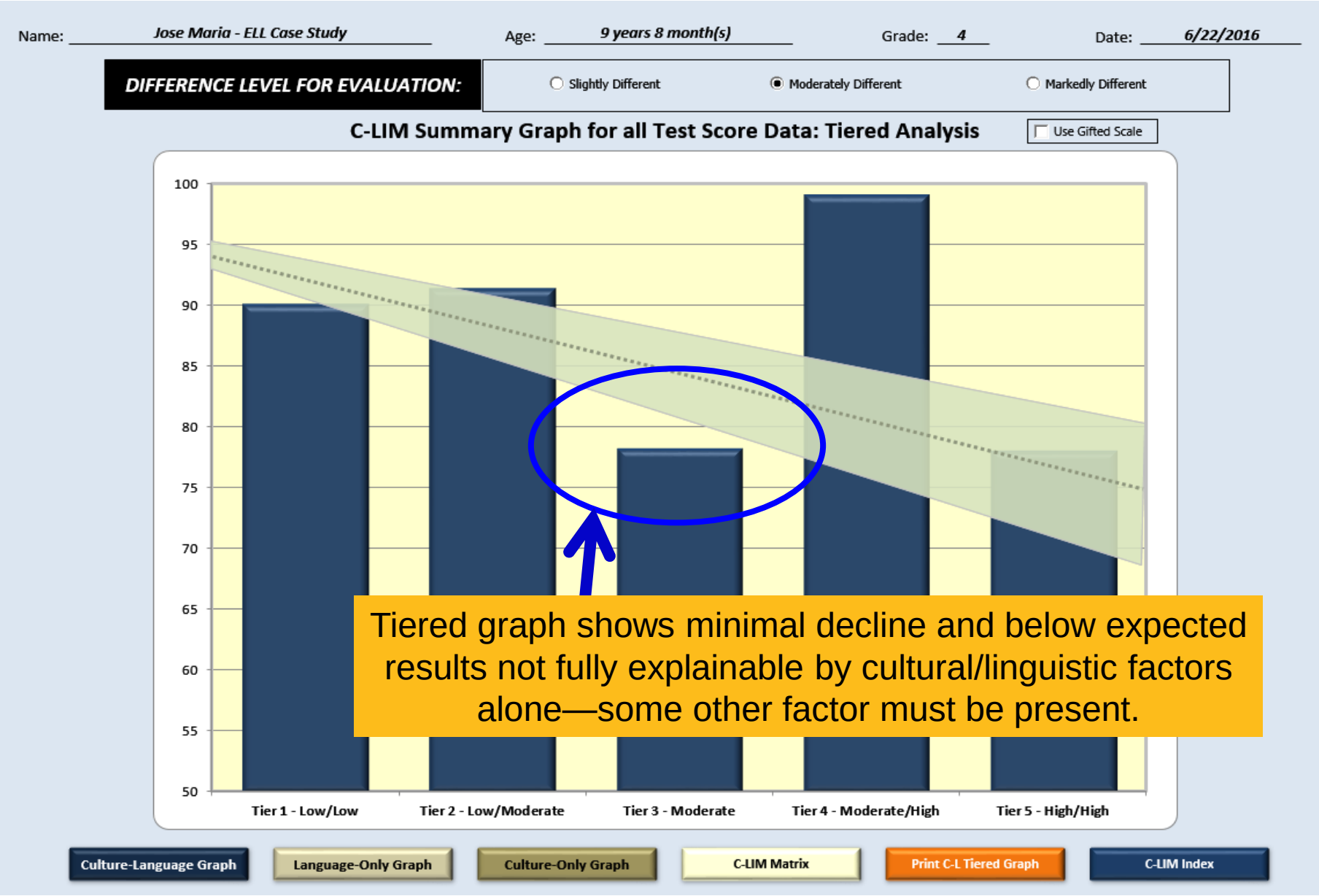
LT Storage/Retrieval	*77*
Story Recall	79
Visual-Auditory Learning	75

SLD Identification with an English Learner: A Case Study

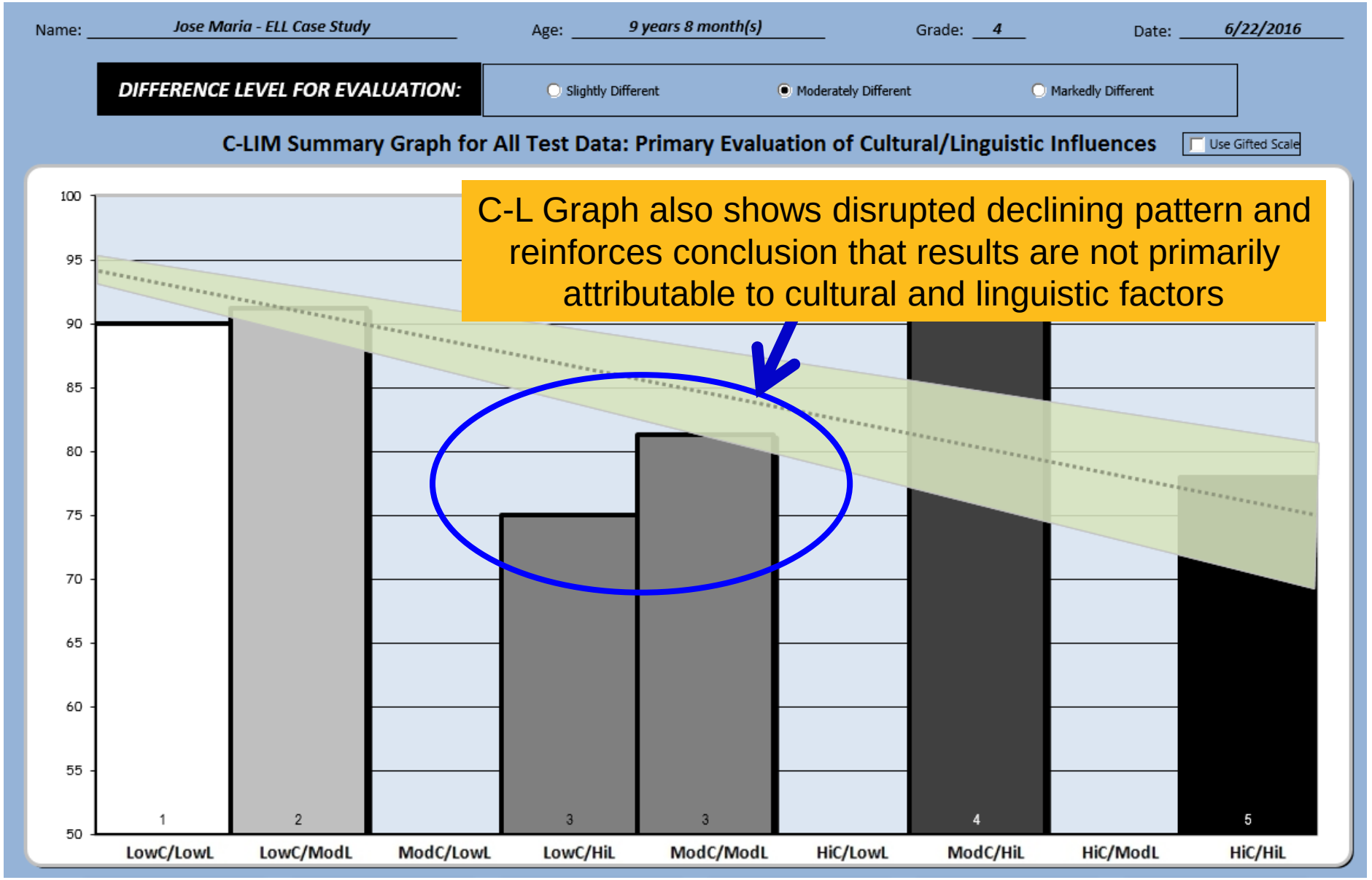


C-LIM is used to interpret pattern of results with respect to whether scores were primarily influenced but cultural/linguistic factors or not.

SLD Identification with an English Learner: A Case Study



SLD Identification with an English Learner: A Case Study



SLD Identification with an English Learner: A Case Study

	A general, overall pattern of decline exists?	All scores within or above the expected range?	All scores within or above the average range?	Degree of influence of cultural and linguistic factors	Likelihood that test scores are valid indicators of ability?
				Primary	Unlikely
				Contributory	Possibly*
Condition C	No	Yes	Yes	Minimal	Likely
Condition D	No	No	No	Minimal	Possibly*

Conclusion from evaluation of score pattern in the C-LIM is consistent with Condition D which indicates lack of general, overall pattern of decline and at least one cell aggregate below the expected range. Therefore, results are only minimally affected by cultural and linguistic issues and are possibly valid—further validation is now required.

**Final determination regarding the validity of test scores that are below the expected and average ranges requires additional data and information, particularly results from native language evaluation, qualitative evaluation and analysis, and data from a strong pre-referral process (e.g., progress monitoring data).*

SLD Identification with an English Learner: A Case Study

Step 1. Evaluate construct validity in all areas in English (exclusion of cultural/linguistic factors)

- *Test in English first and use C-LIM to evaluate scores. If all scores indicate strengths (average or higher) a disability is not likely and no further testing is necessary. If any scores suggest weaknesses, continue evaluation.*

Step 2. Re-evaluate construct validity in areas of weakness in native language (cross-linguistic evidence)

- *If some scores from testing in English indicate weaknesses, re-test those areas in the native language to support them as areas of true weakness*

Step 3. Cross-validate L1 and L2 test scores with contextual factors and data (ecological validity for disability)

- *Use all other case data and information to serve as the context by which to evaluate the test scores and ensure ecological validity to conclusions*

SLD Identification with an English Learner: A Case Study

WISC-V/WJ IV/WIAT-III XBA DATA FOR Jose Maria
DOE: 6/22/2016 DOB: 10/4/2006 Grade: 4

WECHSLER INTELLIGENCE SCALE FOR CHILDREN-V

Verbal Comprehension Index	*76*	*Fluid Reasoning Index*	*82*	*Visual-Spatial Index*	*95*
Similarities	5	Matrix Reasoning	7	Block Design	9
Vocabulary	6	Figure Weights	7	Visual Puzzles	9

Working Memory Index	*79*	*Processing Speed Index*	*94*
Digit Span	5	Coding	9
Picture Span	7	Symbol Search	8

WECHSLER INDIVIDUAL ACHIEVEMENT TEST-III

Basic Reading	*94*	*Reading Comprehension*	*76*	*Written Expression*	*92*
Word Reading	92	Reading Comprehension	76	Spelling	100
Pseudoword Decoding	98	Oral Reading Fluency	80	Sentence Composition	86
				Essay Composition	93

WOODCOCK JOHNSON-IV TESTS OF COGNITIVE ABILITY

Auditory Processing	*91*	*LT Storage/Retrieval*	*77*
Phonological Processing	99	Story Recall	79
Nonword Repetition	84	Visual-Auditory Learning	75

Only areas of possible deficit need to be re-evaluated in the native language (e.g., via use of native language tests, interpreters/translators, etc.). Scores that are average or better do not need to be re-evaluated.

SLD Identification with an English Learner: A Case Study

WISC-V/WJ IV/WIAT-III XBA DATA FOR Jose Maria
DOE: 6/22/2016 DOB: 10/4/2006 Grade: 4

WECHSLER INTELLIEGENCE SCALE FOR CHILDREN-V

Verbal Comprehension Index	*76*
Similarities	5
Vocabulary	6

Fluid Reasoning Index	*82*
Matrix Reasoning	7
Figure Weights	7

Visual-Spatial Index	*95*
Block Design	9
Visual Puzzles	9

Working Memory Index	*79*
Digit Span	5
Picture Span	7

Processing Speed Index	*94*
Coding	
Symbol Search	

WECHSLER INDIVIDUAL ACHIEVEMENT TEST-III

Basic Reading	*94*
Word Reading	92
Pseudoword Decoding	98

Reading Comprehension	
Reading Comprehension	
Oral Reading Fluency	

WOODCOCK JOHNSON-IV TESTS OF COGNITIVE ABILITY

Auditory Processing	*91*
Phonological Processing	99
Nonword Repetition	84

LT Storage/Retrieval	
Story Recall	
Visual-Auditory Learning	75

There are four possible areas of cognitive weakness that may suggest deficits related to the reported academic difficulties as well as three areas of strength. However, because these tests are not designed for English learners, for the areas of suspected weakness it is necessary to generate additional information and data to cross-linguistically confirm that they are true deficits. Strengths do not support disability identification and therefore do not require any further validation.

SLD Identification with an English Learner: A Case Study

WISC-V/WJ IV/WIAT-III XBA DATA FOR Jose Maria
DOE: 6/22/2016 DOB: 10/4/2006 Grade: 4

WECHSLER INTELLIEGENCE SCALE FOR CHILDREN-V

Verbal Comprehension Index	*76*
Similarities	5
Vocabulary	6

Fluid Reasoning Index	*82*
Matrix Reasoning	7
Figure Weights	7

Visual-Spatial Index	*95*
Block Design	9
Visual Puzzles	9

Working Memory Index	*79*
Digit Span	5
Picture Span	7

Processing Speed Index	*94*
Coding	9
Symbol Search	8

WECHSLER INDIVIDUAL ACHIEVEMENT TEST-III

Basic Reading	*94*
Word Reading	92
Pseudoword Decoding	98

Reading Comprehension	
Reading Comprehension	
Oral Reading Fluency	

WOODCOCK JOHNSON-IV TESTS OF COGNITIVE ABILITY

Auditory Processing	*91*
Phonological Processing	99
Nonword Repetition	84

LT Storage/Retrieval	*77*
Story Recall	79
Visual-Auditory Learning	75

In addition, because Gc itself is "language," it cannot be compared fairly to native English speaker norms to determine whether it is a strength or weakness even when scores are deemed "valid" using the C-LIM. Thus, in the case, additional procedures must be employed to determine whether Gc is actually a true weakness or not and whether it does or does not require re-evaluation.

SLD Identification with an English Learner: A Case Study

Interpretive Problems with Gc Scores with English Learners

Because Gc is, by definition, comprised of cultural knowledge and language development, the influence of these factors cannot be separated from tasks designed to measure them. Thus, unless exposure to English is a controlled variable in a test's norm sample and the sample includes many different languages, *Gc scores for ELLs always remain at risk for inequitable interpretation even when the overall pattern of scores within the C-LIM is determined to be valid.*

For example, a Gc score of 76 would be viewed as "deficient" relative to a norm sample comprised primarily of native English speakers. Moreover, testing in the native language doesn't solve this problem because current native-language tests treat ELs as being all the same (they aren't), as if being behind in English is only temporary (it isn't), as if the country they come from is important (it's not), and as if five years of English learning makes them native English speakers (it doesn't).

Therefore, practitioners must find and rely on a "true peer" comparison group such as that which is formed within the High Culture/High Language cell of the C-LIM to help *ensure that ELLs are not unfairly regarded as having either deficient Gc ability or significantly lower overall cognitive ability*—conditions that may simultaneously decrease identification of SLD and increase suspicion of ID and speech impairment.

SLD Identification with an English Learner: A Case Study

Determining if and when to re-test Gc via the C-LIM

Re-evaluation of suspected areas of weakness is necessary to provide cross-linguistic confirmation of potential deficits in functioning. A disability cannot be identified in an English learner if the observed difficulties occur only in one language. Even then, deficits that are identified in both languages are not definitive evidence of dysfunction and evaluation of expectations for native language performance is as relevant for native language evaluation as it is for evaluation in English.

Because of the nature of Gc, it should be treated slightly differently when it comes to re-evaluation as compared to other cognitive abilities. The following guidelines from the best practice recommendations apply specifically to Gc:

- **Review results from testing in English and identify domains of suspected weakness or difficulty:*
    a. *For Gc only, evaluate weakness according to high/high cell in C-LIM or in context of other data and information*
- **For Gc only:*
    a. *If high/high cell in C-LIM is within/above expected range, consider Gc a strength and assume it is at least average (re-testing is not necessary)*
    b. *If high/high cell in C-LIM is below expected range, re-testing of Gc in the native language is recommended*
- *For Gc only, scores obtained in the native language should only be interpreted relative to developmental and educational experiences of the examinee in the native language and only as compared to others with similar developmental experiences in the native language.*

It is important that the actual, obtained Gc score, regardless of magnitude, be reported when required, albeit with appropriate nondiscriminatory assignment of meaning, and that it be used for the purposes of instructional planning and educational intervention.

****If Gc is evaluated with the Ortiz PVAT, use the actual score obtained from the English Learner norms (NOT the English Speaker norms) to determine if it is an area of weakness. If the score indicates a weakness, it should then be further re-evaluated in the native language.***

SLD Identification with an English Learner: A Case Study

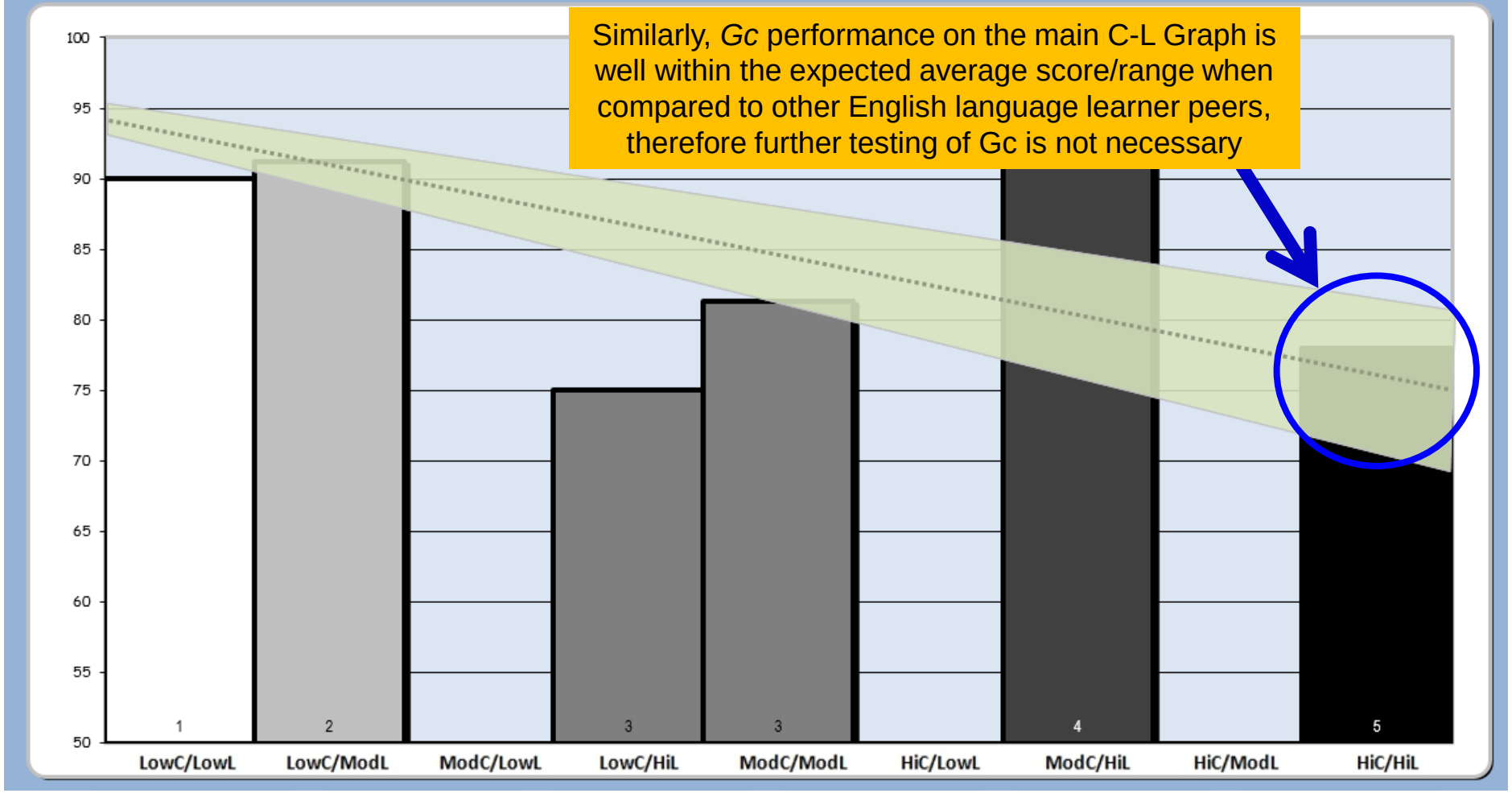
Gc performance on the C-LIM Summary Graph is well within the expected average score/range when compared to other English language learner peers, therefore further testing of Gc is not necessary

SLD Identification with an English Learner: A Case Study

Name: Jose Maria - ELL Case Study Age: 9 years 8 month(s) Grade: 4 Date: 6/22/2016

DIFFERENCE LEVEL FOR EVALUATION: ○ Slightly Different ◉ Moderately Different ○ Markedly Different

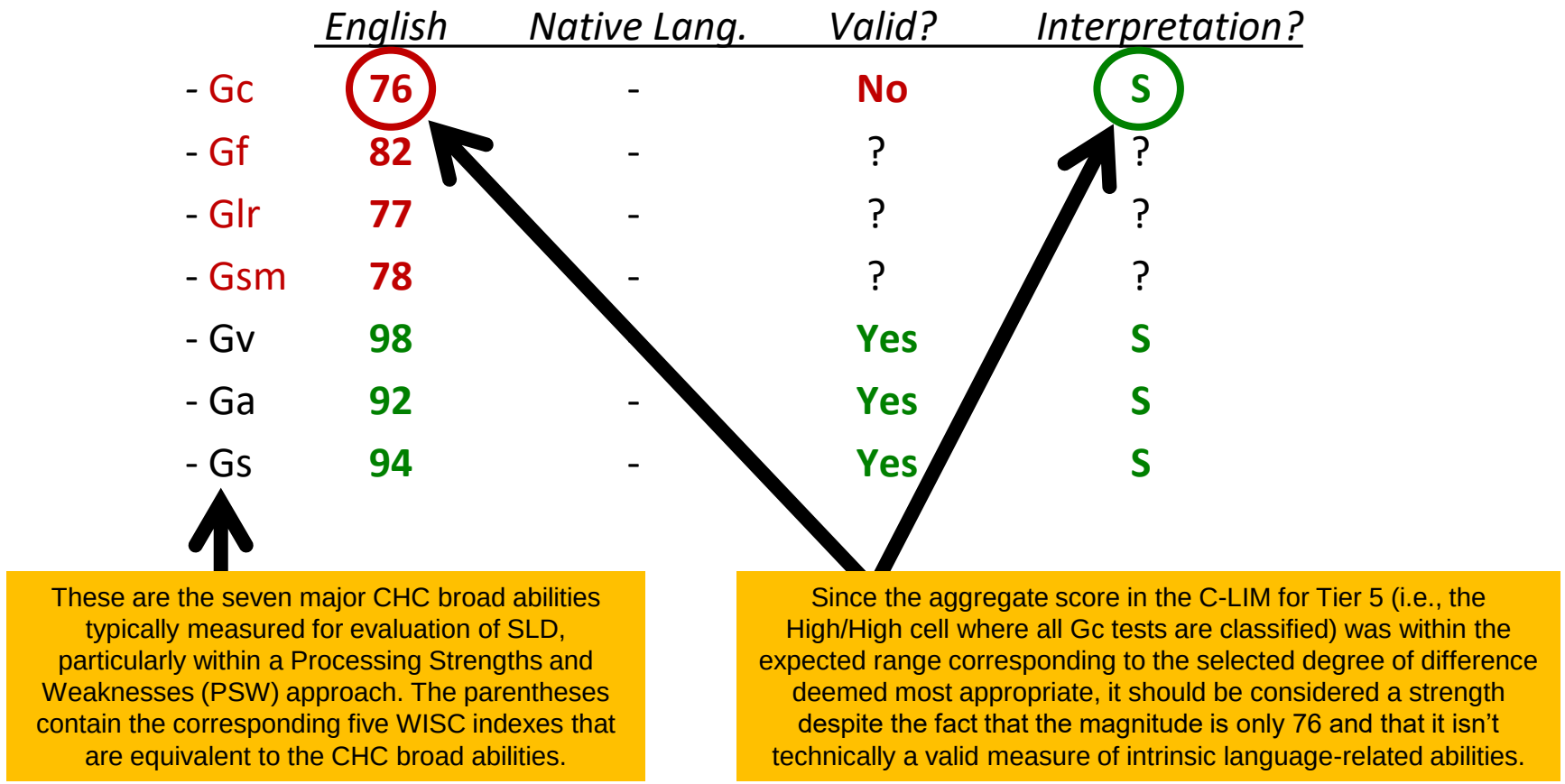
C-LIM Summary Graph for All Test Data: Primary Evaluation of Cultural/Linguistic Influences ☐ Use Gifted Scale



SLD Identification with an English Learner: A Case Study

Interpretive Problems with Gc Scores with English Learners

Although the C-LIM helped determine that Gc is NOT an area of weakness, further evaluation and interpretation is complicated because of the low magnitude of the score (i.e., SS=76). Other corrections are necessary to prevent discriminatory decisions, particularly in evaluation of SLD or SLI. However, use of the Ortiz PVAT provides a simple and more direct solution to all of these problems.



SLD Identification with an English Learner: A Case Study

Resolving Problems with Gc Scores for ELs: The Ortiz PVAT

Clearly, the preceding procedures necessary to address validity issues related to the measurement of Gc and language/culture-related abilities are complicated, somewhat cumbersome, and not very efficient. It may also leave the practitioner in the unenviable position of having to defend a very low score (SS=76) as being technically invalid, but still considered to be an area of processing "strength."

This one issue, more than any other, best highlights the shortcomings of today's tests relative to their failure to provide a true peer comparison group for English learners that would alleviate all of the extra work and potential confusion. There simply is no substitute for being able to make fair and equitable interpretations than comparison to peers with similar developmental experiences.

That said, there is in fact an easier way to do all of this. In response to the many difficulties posed by these issues, *a new test has been developed with dual-norm samples, including one specifically for English learners that yields valid Gc scores for English learners of any language background and level of English exposure—and that test is the* ***Ortiz PVAT*****.**

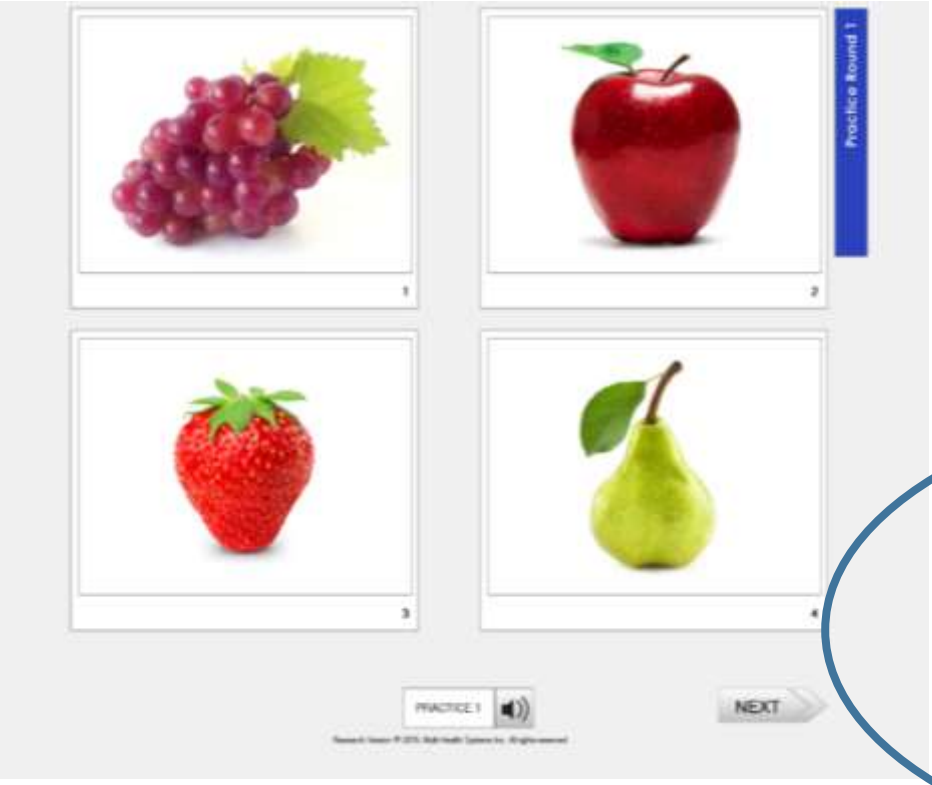
The Ortiz PVAT – A new direction in tests and testing.

Clinical and Educational Applications of the Ortiz PVAT

- Diagnostic evaluation – provides “true peer” comparisons for evaluating language-related disabilities/disorders in both English Speakers and Learners
- Intervention/treatment – provides data and specific recommendations for language-based intervention keyed directly to performance relative to peers
- Instructional guidance – provides data and specific teaching and instructional recommendations based on performance relative to grade-level expectations
- Progress monitoring – provides data for documenting progress across short intervals to evaluate success of instruction and intervention efforts
- Growth – provides data and a specific Index capable of documenting actual growth in vocabulary/language acquisition across short and long intervals

The Ortiz PVAT – A new direction in tests and testing.

The Ortiz PVAT is a computer-based assessment that measures vocabulary acquisition in children and youths aged 2.6 to 22 using English language words and irrespective of the native/heritage language.



Perhaps the most unique feature of the Ortiz PVAT is:

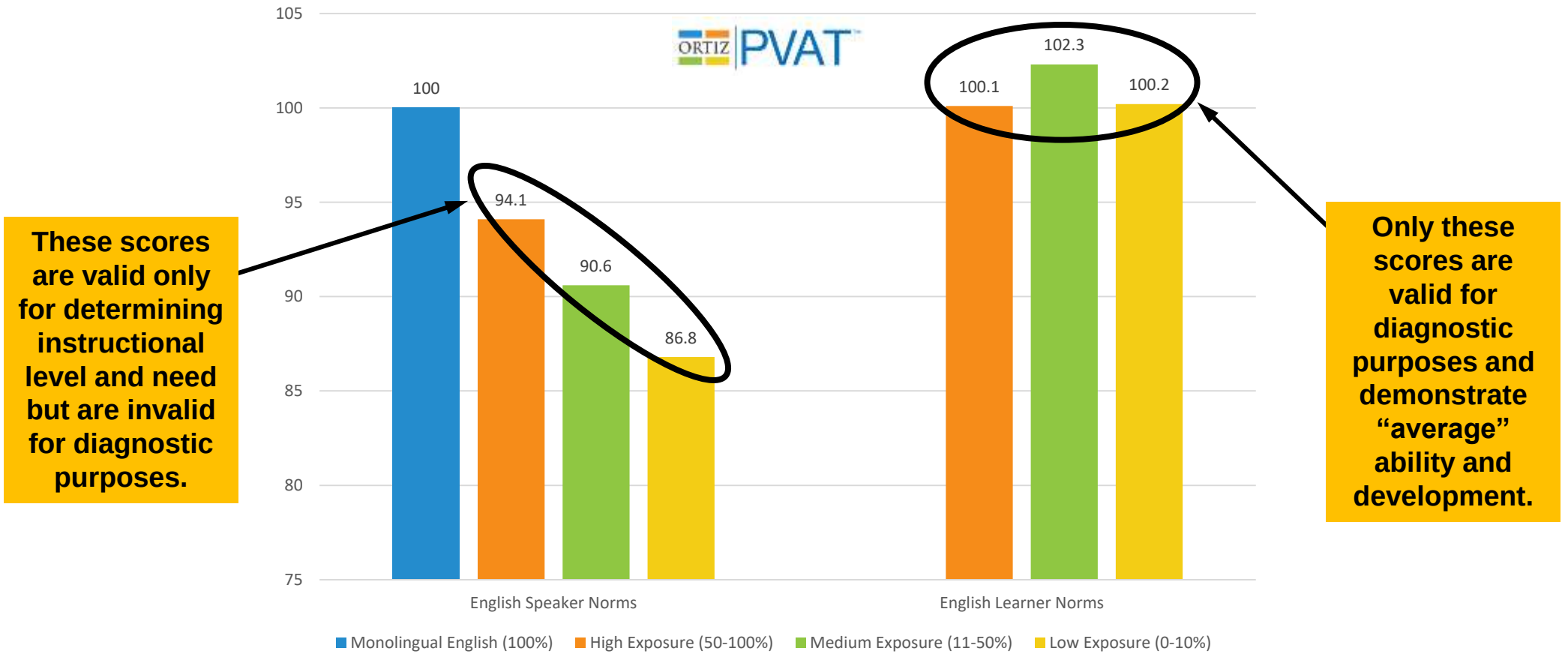
A “dual-norming” structure with distinct norms for English-speakers and English learners.

The EL norms are based on ELs from various language backgrounds and specifically control for amount of English exposure.

MHS

Performance of English Learners Based on Comparison to English Learner vs. English Speaker Norm Samples

Developmental Language/Exposure-based Comparison Provides Validity and Fairness for ELs



Fairness Across All English Learners:

No evidence of bias regarding native language

Assumption: English language acquisition is an invariant process, irrespective of the native language

Form	Language Spoken	*N*	*M*	*SD*	*F* (*df*)	*p*	Pairwise Comparisons ($p < .01$)	Partial η^2
Form A	Spanish & Spanish Creole	872	101.5	15.5	1.63 (3, 1183)	.181	*ns*	.004
	Indo-European Languages	161	99.4	15.7				
	Asian & Pacific Islander Languages	129	98.8	15.4				
	All Other Languages	28	99.9	15.4				
Form B	Spanish & Spanish Creole	872	101.7	15.5	1.52 (3, 1183)	.208	*ns*	.004
	Indo-European Languages	161	99.8	15.7				
	Asian & Pacific Islander Languages	129	99.0	15.4				
	All Other Languages	28	99.9	15.4				

SLD Identification with an English Learner: A Case Study

WISC-V/WJ IV/WIAT-III XBA DATA FOR Jose Maria
DOE: 6/22/2016 DOB: 10/4/2006 Grade: 4

Although we are adding the Ortiz PVAT at this point in the evaluation, it would have been easiest to simply include it as a standard part of any battery particularly because it can be administered to any individual to generate a valid Gc score, and in the case of ELs, it will also address the Gc problem that will always exist and provide that information in an interpretive summary report.

WECHSLER INTELLIEGENCE SCALE FOR CHILDREN-V

Verbal Comprehension Index	*76*
Similarities	5
Vocabulary	6

Fluid Reasoning Index	*82*
Matrix Reasoning	7
Figure Weights	7

Working Memory Index	*79*
Digit Span	5
Picture Span	7

Processing Speed Index	*94*
Coding	9
Symbol Search	8

WECHSLER INDIVIDUAL ACHIEVEMENT TEST-III

Basic Reading	*94*
Word Reading	92
Pseudoword Decoding	98

Reading Comprehension	*76*
Reading Comprehension	76
Oral Reading Fluency	80

Written Expression	*92*
Spelling	100
Sentence Composition	86
Essay Composition	93

WOODCOCK JOHNSON-IV TESTS OF COGNITIVE ABILITY

Auditory Processing	*91*
Phonological Processing	99
Nonword Repetition	84

LT Storage/Retrieval	*77*
Story Recall	79
Visual-Auditory Learning	75

Ortiz PVAT (EL Norms)	*93*

SLD Identification with an English Learner: A Case Study

Avoiding Interpretive Problems by Use of the Ortiz PVAT

Derivation of an Ortiz PVAT score using the English learner norms eliminates the Gc problem completely. The Ortiz PVAT score simply replaces any Gc/language-related/verbal ability score because it was derived precisely on "true peers" and therefore inherently valid in terms of both meaning/classification and actual magnitude (e.g., 90 - 109 = average).

	English	*Spanish*	*Valid?*	*Interpretation?*
- Gc	**76**	-	**No**	?
- Gf	**82**	-	?	?
- Glr	**77**	-	?	?
- Gsm	**78**	-	?	?
- Gv	**98**	-	**Yes**	**S**
- Ga	**92**	-	**Yes**	**S**
- Gs	**94**	-	**Yes**	**S**
- Gc (Ortiz PVAT)	**93**	-	**Yes**	**S**

Use of the Ortiz PVAT requires no native language confirmation since the score is derived from norms that control for amount of exposure to English and is based on a true peer comparison group for both English speakers and English learners. Therefore, it is valid and may be interpreted directly as a strength or weakness without requiring any further cross-linguistic validation. It also eliminates the potential confusion and difficulty in having to explain why a low score (e.g. 76) is a strength, not a weakness.

Nondiscriminatory Interpretation of Test Scores: A Case Study

Determining if and when to re-evaluate all other (non-Gc) abilities

Because cultural knowledge and language ability are not the primary focus in measurement of other abilities, the influence of cultural/linguistic factors can be determined via the C-LIM and scores below the expected range of performance may well be deemed to be the result of factors other than cultural knowledge or language ability. Thus, there is no limitation requiring comparison of performance to a true ELL peer group as there is with Gc. Thus, use of a test's norms and the attendant standard classification scheme is appropriate for determining areas of suspected weakness using tests administered in English for abilities other than Gc.

However, to establish validity for a low score obtained from testing in English with an ELL, native language evaluation is required. The following guidelines from the best practice recommendations apply to all abilities, including Gc—when Gc has been determined to be a weakness because it falls below the expected range of difference in the C-LIM:*

- *Review results from testing in English and identify domains of suspected weakness or difficulty:*
    - *a. For all abilities, except Gc, evaluate weakness using standard classifications (e.g., SS < 90)*
- *Re-test all domains of suspected weakness, including Gc when it is not within the expected range of difference in the C-LIM* using native language tests*
- *Administer native language tests or conduct re-testing using one of the following methods:*
    - *a. Native language test administered in the native language (e.g., WJ III/Bateria III or WISC-IV/WISC-IV Spanish)*
    - *b. Native language test administered via assistance of a trained interpreter*
    - *c. English language test translated and administered via assistance of a trained interpreter*
- *Administer tests in manner necessary to ensure full comprehension including use of any modifications and alterations necessary to reduce barriers to performance, while documenting approach to tasks, errors in responding, and behavior during testing, and analyze scores both quantitatively and qualitatively to confirm and validate areas as true weaknesses*

Or, if Gc was evaluated with the Ortiz PVAT, the actual score when compared to the English Learner norms (NOT the English Speaker norms) indicates that it is likely an area of weakness.

SLD Identification with an English Learner: A Case Study

Procedures for Follow-up Evaluation in the Native Language

When providing cross-linguistic confirmation of areas of weakness that were found via scores derived from testing in English, it is helpful (but not actually necessary) to generate scores. Qualitative information and data (e.g., process or error analysis, dynamic assessment, task observations, etc.) are equally helpful and useful with respect to confirming areas of weakness.

It is also reasonable to use the exact same tests for follow up evaluation in the native language as were initially used in English language evaluation because, in this case, practice effects are diagnostically helpful in terms of discerning "learning ability" from "learning disability."

Evaluation in the native language can be accomplished in several different ways and will likely depend on the competency of the evaluator and the available resources. Completion of the task may include one or more of the following procedures:

1. *Use of native language tests (if available) administered by a bilingual evaluator*
2. *Use of native language tests (if available) administered by a trained translator*

In the absence of parallel or similar native language tests with which to evaluate the necessary domains, follow up evaluation will need to resort to other procedures for task completion, including:

3. *Use of English language tests translated directly by a bilingual evaluator*
4. *Use of English language tests administered via assistance of trained translator*
5. *Use of informal tasks accompanied by careful observation, error analysis, and other probing with the assistance of a translator for communication.*

SLD Identification with an English Learner: A Case Study

WISC-V/WJ IV/WIAT-III XBA DATA FOR Jose Maria
DOE: 6/22/2016 DOB: 10/4/2006 Grade: 4

WECHSLER INTELLIGENCE SCALE FOR CHILDREN-V

Verbal Comprehension Index	*76*
Similarities	5
Vocabulary	6

Fluid Reasoning Index	*82*
Matrix Reasoning	7
Figure Weights	7

Visual-Spatial Index	*95*
Block Design	9
Visual Puzzles	9

Working Memory Index	*79*
Digit Span	5
Picture Span	7

Processing Speed Index	*94*
Coding	9
Symbol Search	8

WISC IV Spanish **(Gf subtests)**	91
Matrix Reasoning	8
Picture Concepts	9

WISC IV Spanish WMI	*72*
Digit Span	5
Letter-Number Sequencing	4

WOODCOCK JOHNSON-IV TESTS OF COGNITIVE ABILITY

Auditory Processing	*91*
Phonological Processing	99
Nonword Repetition	84

LT Storage/Retrieval	*77*
Story Recall	79
Visual-Auditory Learning	75

Ortiz PVAT	*93*

Bateria III LT Retrieval	*79*
Visual-Auditory Learning	81
Retrieval Fluency	78

Gf, Gsm, and Glr need to be re-tested in the native language to provide additional confirmation that they are true weaknesses. The same or similar tests can be used and scores may be generated but the main purpose is to observe performance qualitatively in the domain to provide cross-linguistic validation of suspected difficulties.

SLD Identification with an English Learner: A Case Study

WISC-V/WJ IV/WIAT-III XBA DATA FOR Jose Maria
DOE: 6/22/2016 DOB: 10/4/2006 Grade: 4

WECHSLER INTELLIEGENCE SCALE FOR CHILDREN-V

Verbal Comprehension Index	*76*
Similarities	5
Vocabulary	6

Fluid Reasoning Index	*82*
Matrix Reasoning	7
Figure Weights	7

Visual-Spatial Index	*95*
Block Design	9
Visual Puzzles	9

Working Memory Index	*79*
Digit Span	5
Picture Span	7

Processing Speed Index	*94*
Coding	9
Symbol Search	8

WISC IV Spanish **(Gf subtests)**	91
Matrix Reasoning	8
Picture Concepts	9

WISC IV Spanish WMI	*72*
Digit Span	5
Letter-Number Sequencing	4

WOODCOCK JOHNSON-IV TESTS OF COGNITIVE ABILITY

Auditory Processing	*91*
Phonological Processing	99
Nonword Repetition	84

LT Storage/Retrieval	*77*
Story Recall	7
Visual-Auditory Learning	75

Ortiz PVAT	*93*

Bateria III LT Retrieval	*79*
Visual-Auditory Learning	81
Retrieval Fluency	78

Results of native language testing for Gf, Gsm, and Glr

Nondiscriminatory Interpretation of Test Scores: A Case Study

Determining which scores are valid and interpretable

Average* or higher scores in testing are unlikely to be due to chance. Thus, when a score obtained from native language testing is found to be in the average range or higher, it serves to effectively invalidate the original low score from testing in English since deficits must exist in both languages. Conversely, if another low score in the same domain is obtained from native language evaluation, it may serve to bolster the validity of the original score obtained in English.

Based on these premises, the following guidelines from the best practice recommendations offer guidance regarding selection and use of the most appropriate and valid score for the purposes of PSW analysis (or any other situation in which the validity of test scores is central or relevant):

- *For all domains, including Gc, if a score obtained in the native language suggests a domain is a strength (SS ≥ 90), it serves to invalidate/disconfirm the corresponding weakness score obtained in English—thus, report, use, and interpret the domain score obtained in the native language*
- *For all domains, except Gc, if a score obtained in the native language also suggests weakness in the same domain (SS < 90), it serves to validate/confirm the corresponding weakness score obtained in English—thus, report, use, and interpret the original domain score obtained in English*
- *For Gc only, if a score obtained in the native language also suggests weakness in Gc (SS < 90), it may serve to validate/confirm the corresponding weakness score obtained in English but only if low performance in Gc cannot be attributed to factors related to a lack or interruption of native language instruction and education, low family SES, or other lack of opportunity to learn—thus, in the absence of such mitigating factors, report, use, and interpret the domain score obtained in English*

**Although "average or higher" (e.g., SS≥90) is used as a recommended cutoff for supporting the validity of test scores, use of a lower standard (e.g., SS≥85) may also represent a reasonable standard for practice since it is based on performance that can be categorized as being within normal limits.*

A Recommended Best Practice Approach for Using Tests with ELLs

ADDRESSING VALIDITY AND INTERPRETATION OF SCORES GENERATED IN TWO LANGUAGES

	Original score when tested in English	Follow up score when tested in native language	Most appropriate and valid score for use in PSW analysis		Rationale for Use as Strength or Weakness in PSW Analysis
			Original Score (in English)	Follow Up Score (in native lang)	
For ALL domains*	**SS ≥ 90**	**n/a**	✓		**Strength—scores in or above the average range (or even WNL) are unlikely to occur by chance and very likely to be valid thus re-evaluation in the native language is unnecessary**
For ALL domains (and when Gc is below expected range in C-LIM)	**SS < 90**	**SS ≥ 90**		✓	**Strength—because a deficit cannot exist in one language only, the original score from testing in English is invalidated and should be replaced by the follow up average score which is likely to be valid**
For ALL domains (and when Gc is below expected range in C-LIM)	**SS < 90**	**SS < 90**	✓		**Weakness—low scores in both languages suggest a true deficit but additional, convergent and consistent ecological evidence is required to substantiate scores as deficits**
For Gc Only (and when Gc is within the expected range in C-LIM)	**SS < 90**	**n/a**	✓		**Strength—Gc can only be compared fairly to other ELLs, thus its position within the expected range in the C-LIM should be considered to be average and native language testing may not be necessary unless there is reason to believe it may be informative**

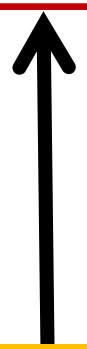
**Although this table uses "average or higher" (e.g., SS≥90) as a recommended cutoff for supporting the validity of test scores, use of a lower standard (e.g., SS≥85) may also represent a reasonable standard for practice since it is based on performance that can be categorized as being within normal limits.*

SLD Identification with an English Learner: A Case Study

Determining which scores are valid and interpretable

Derivation of an Ortiz PVAT score using the English learner norms eliminates the Gc problem completely. The Ortiz PVAT score simply replaces any Gc/language-related/verbal ability score because it was derived precisely on EL "true peers" and therefore inherently valid in terms of both meaning/classification and actual magnitude (e.g., 90 - 109 = average).

	English	*Spanish*	*Valid?*	*Interpretation?*
- Gc	**76**	-	76 - **No**	-
- Gf	**(82)**	**91**	**91 - Yes**	**S**
- Glr	**77**	**(79)**	**77 - Yes**	**W**
- Gsm	**78**	**(72)**	**78 - Yes**	**W**
- Gv	**98**	-	**Yes**	**S**
- Ga	**92**	-	**Yes**	**S**
- Gs	**94**	-	**Yes**	**S**
- Gc (Ortiz PVAT)	93	-	**Yes**	**S**



Additional native language investigation of areas of weakness noted in scores derived from testing in English (with the exception of the score from the Ortiz PVAT), resulted in an average Gf score that invalidated the original Gf score, and two below average scores that simply cross-linguistically confirmed Glr and Gsm as areas of weakness as indicated by the test scores in English.

SLD Identification with an English Learner: A Case Study

Step 1. Evaluate construct validity in all areas in English (exclusion of cultural/linguistic factors)

- *Test in English first and use C-LIM to evaluate scores. If all scores indicate strengths (average or higher) a disability is not likely and no further testing is necessary. If any scores suggest weaknesses, continue evaluation.*

Step 2. Re-evaluate construct validity in areas of weakness in native language (cross-linguistic evidence)

- *If some scores from testing in English indicate weaknesses, re-test those areas in the native language to support them as areas of true weakness*

Step 3. Cross-validate L1 and L2 test scores with contextual factors and data (ecological validity for disability)

- *Use all other case data and information to serve as the context by which to evaluate the test scores and ensure ecological validity to conclusions*

SLD Identification with an English Learner: A Case Study

The Importance of Converging Evidence in Establishing Validity

Validity is based on an accumulation of evidence. The evaluation approach described herein is designed to assist in generating test scores that may be interpreted as valid indicators of an individual's abilities. Embedded in the broader framework are two basic forms of evidence that bolster the validity of obtained test scores by using expectations of test performance that are grounded in research on individuals of comparable cultural and linguistic backgrounds and the extent to which their development differs from the individuals on whom the tests were normed. Validity is thus inferred by:

1. Test scores from evaluation in English that have been subjected to systematic analysis of the influence of cultural and linguistic variables where such factors have been found to be either minimal or contributory but not primary factors in test performance;

2. Test scores or qualitative data regarding evaluation of weak areas in the native language that either further confirm suspected areas of deficit as being true or dis-confirm suspected areas of deficit due to evidence of average or higher performance.

To these two forms of evidence, a third should be added to fully support conclusions and interpretation of the obtained test scores:

3. Ecological and contextual evidence regarding consistency of the test scores with ecological data and information on developmental influences (e.g., L1 and L2 exposure, language of instruction, socio-economic status, parental education level, etc.) and convergence of patterns of performance with other case data (e.g., progress monitoring data, pre-referral concerns, work samples, observations, school records, teacher/parent reports, grades, interviews, observations, etc.).

Only when all three forms of evidence are seen to converge can there be sufficient confidence in the use and interpretation of test scores obtained in an evaluation of English learners.

SLD Identification with an English Learner: A Case Study

Corroborating test scores with additional, converging evidence

	English	*Spanish*	*Valid?*	*Interpretation?*
- Gc	**76**	-	**No**	-
- Gf	**(82)**	**91**	**Yes**	**S**
- Glr	**77**	**(79)**	**Yes**	**W**
- Gsm	**78**	**(72)**	**Yes**	**W**
- Gv	**98**	-	**Yes**	**S**
- Ga	**92**	-	**Yes**	**S**
- Gs	**94**	-	**Yes**	**S**
- Gc (Ortiz PVAT)	**93**	-	**Yes**	**S**

The areas of weakness identified in this case are in the domains of Glr and Gsm. Additional converging evidence that these are true weaknesses comes from both the additional native language evaluation results and corroborating information from pre- or post-referral interviews, record reviews, observations, work samples, etc., all of which are consistent in demonstrating that the individual has problems on tasks that require long-term memory or short-term memory skills (e.g., inconsistent learning, lack of expected fluency and automaticity, easily forgetting things, inability to follow multi-step directions, errors in procedural steps in math calculations, difficulty remembering what was just read, etc.).

SLD Identification with an English Learner: A Case Study

Sample Validity Statement for ELL Evaluations

Statement 2. Evaluations of Suspected Learning Disability - Valid Results

The following sample validity statement may be used in cases where a clear declining pattern is NOT evident, that is, there is no primary effect of culture and language thus the results ARE valid and there may be a disability.

Because the student is not a native English speaker, it is necessary to establish the validity of the results obtained from testing to ensure that they are accurate estimates of ability or knowledge and not the manifestation of cultural or linguistic differences. To this end, a systematic evaluation of the possible effects of a relative lack of opportunity for the acquisition of acculturative knowledge and English proficiency was carried out via use of the Culture-Language Interpretive Matrix (C-LIM).

A careful review of the student's test data, as entered into the C-LIM, revealed either no overall pattern of decline or a partial pattern of decline combined with performance in one or more area that was below the range that would be expected of other individuals with similar cultural and linguistic backgrounds. This pattern of test performance suggests that cultural and linguistic factors were either ***minimal*** (no evident decline) or ***contributory*** (some decline) influences on the measured test performance but can not account for the entirety of the results. Accordingly, the test results were not considered to be due primarily to the influence of cultural and linguistic factors but still required additional information to fully establish their validity. Evidence to further support the validity of the obtained results was provided by converging sources of information including results from native language evaluation, progress-monitoring data, qualitative analysis, and authentic assessment methods. In addition, other extraneous factors that might account for the observed pattern (for example, lack of motivation, fatigue, incorrect administration/scoring, emotional/behavioral problems) were also evaluated and excluded. Taken together, the reported test results were deemed likely to be valid, interpretable, and to be reliable estimates of the student's actual ability or knowledge. However, equitable interpretation of Gc (cultural knowledge and language development), required comparison relative to other English learners with comparable linguistic development and educational experiences which was accomplished via examination of the magnitude of the high culture/high language cell in the C-LIM and whether it was within the selected range of difference. Consequently, the academic difficulties observed in classroom performance and which prompted this evaluation are not likely to attributable primarily to the process of normal second language and acculturative knowledge acquisition.

In summary, the observed pattern of the student's test results is not consistent with performance that is typical of non-disabled, culturally and linguistically diverse individuals who are of average ability or higher. Therefore, it can be reasonably concluded that the test data evaluated with the C-LIM are likely to be valid, are supported by additional converging data, and suggest that that the student's test performance can be used to support the presence of a learning disability or other cognitive-based disorder.

The statement above is the one most appropriate for this case where a) the evaluation focused on suspected SLD; and b) where it was determined that the obtained test results were NOT influenced primarily by cultural and linguistic factors, albeit they remained contributory. Thus, the test results (except for Gc) could be considered valid estimates of the abilities that were measured. In addition, native language testing was conducted to further support cognitive test score validity. This statement (and three others) have been placed in the public domain and may be freely copied, modified, and distributed for non-profit purposes without the need to secure permission.

Culture-Language Interpretive Matrix: The Importance of the Context of Difference

Subtests	Standard Score	Confidence Interval (95% Band)	Descriptions
Verbal Comprehension	64	56 – 72	Very Low
Visual-Auditory Learning	88	76 – 100	Low Average
Spatial Relations	98	91 – 107	Average
Sound Blending	75	64 – 87	Low
Concept Formation	70	62 – 78	Low
Visual Matching	86	76 – 97	Low Average
Numbers Reversed	80	67 – 93	Low
Incomplete Words	78	65 – 91	Low
Auditory Working Memory	85	76 – 94	Low Average
Analysis-Synthesis	78	66 – 90	Low
Auditory Attention	81	67 – 95	Low
Decision Speed	72	63 – 81	Low
Retrieval Fluency	82	69 – 95	Low
General Information	69	60 – 78	Very Low

Culture-Language Interpretive Matrix: The Importance of the Context of Difference

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

Name: _____ Age: _____ Grade: _____ CLEAR DATA SAVE DATA

DEGREE OF LINGUISTIC DEMAND

DEGREE OF CULTURAL LOADING

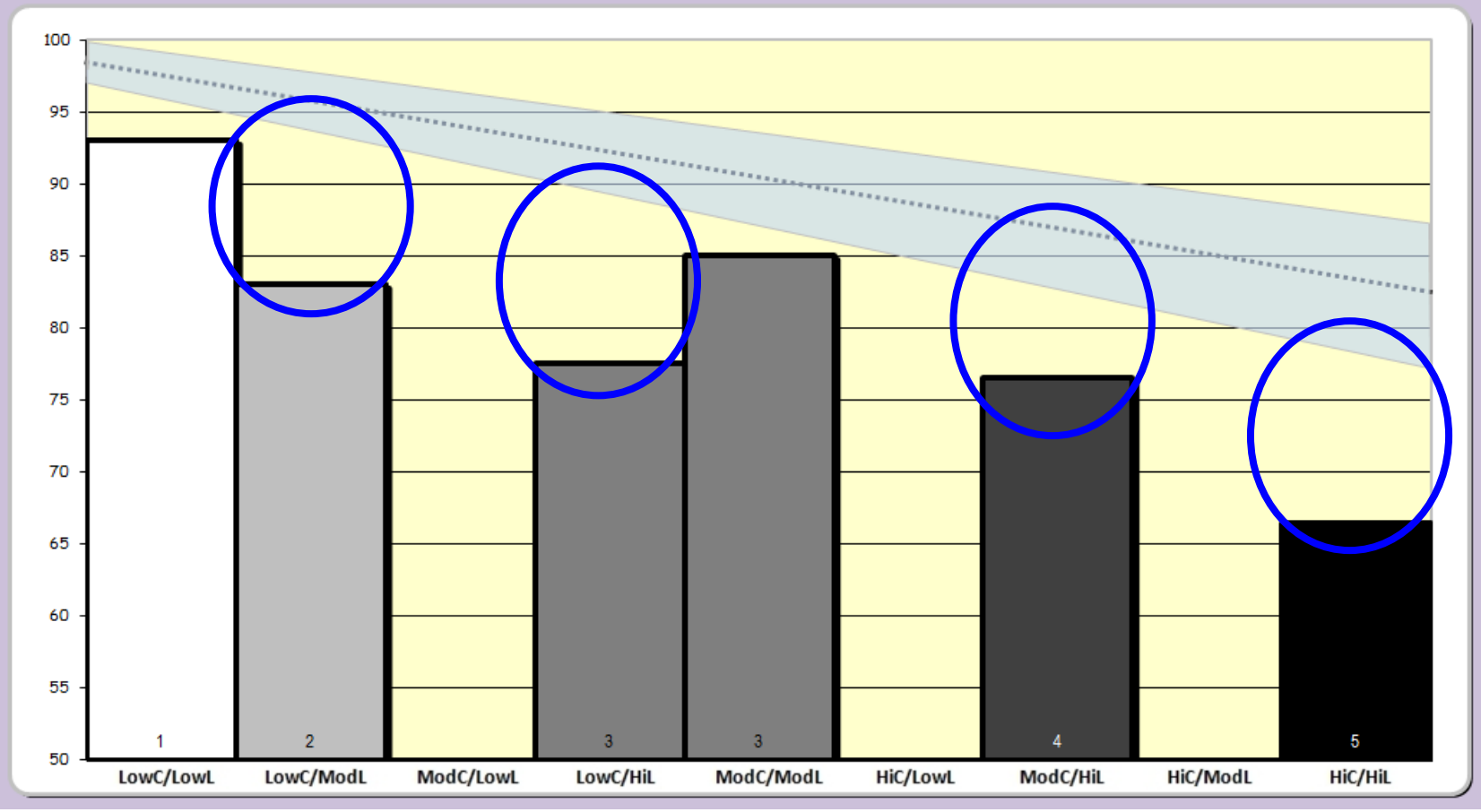
	LOW		MODERATE		HIGH	
LOW		Score		Score		Score
	WJ III Spatial Relations	93 (93)	WJ III Numbers Reversed	80 (80)	WJ III Analysis-Synthesis	78 (78)
			WJ III Visual Matching	86 (86)	WJ III Auditory Working Memory	85 (85)
					WJ III Concept Formation	70 (70)
	Cell Average =	93	Cell Average =	83	Cell Average =	78
MODERATE		Score		Score		Score
	WJ III Pair Cancellation		WJ III Delayed Recall: Visual Auditory Learning		WJ III Auditory Attention	81 (81)
	WJ III Picture Recognition		WJ III Rapid Picture Naming		WJ III Decision Speed	72 (72)
	WJ III Planning		WJ III Retrieval Fluency	82 (82)	WJ III Incomplete Words	78 (78)
			WJ III Visual Auditory Learning	88 (88)	WJ III Memory for Words	
					WJ III Sound Blending	75 (75)
	Cell Average =		Cell Average =	85	Cell Average =	77
HIGH		Score		Score		Score
					WJ III General Information	69 (69)
					WJ III Verbal Comprehension	64 (64)
	Cell Average =		Cell Average =		Cell Average =	67

Culture-Language Interpretive Matrix: The Importance of Difference

Name: _____ Age: _____ Grade: _____

DIFFERENCE LEVEL FOR EVALUATION: (•) Slightly Different ( ) Moderately Different ( ) Markedly Different

XBA C-LIM Graph for WJ III NU COG: Primary Evaluation of Cultural and Linguistic Influences

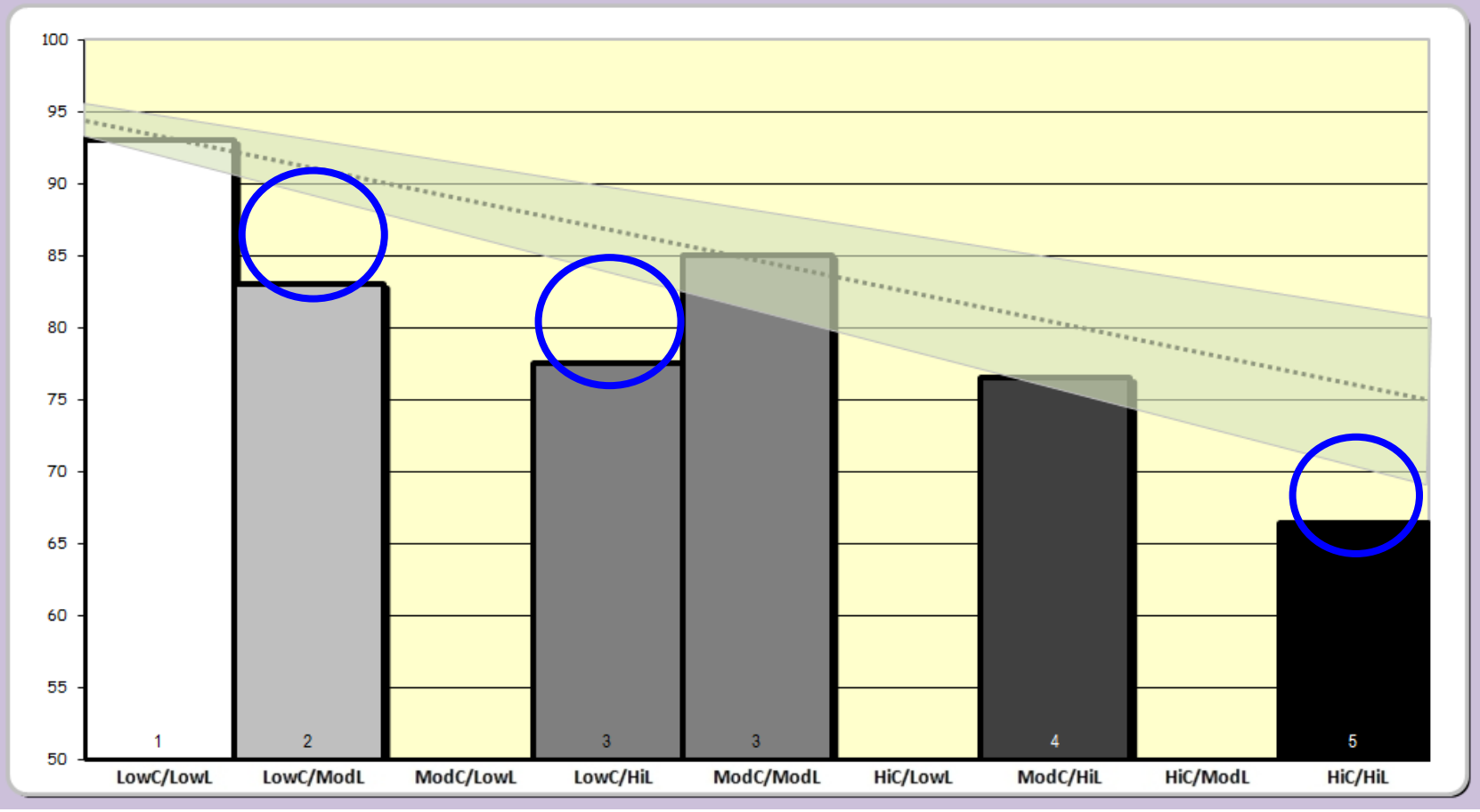


Culture-Language Interpretive Matrix: The Importance of the Context of Difference

Name: _____ Age: _____ Grade: _____

DIFFERENCE LEVEL FOR EVALUATION: ○ Slightly Different ◉ Moderately Different ○ Markedly Different

XBA C-LIM Graph for WJ III NU COG: Primary Evaluation of Cultural and Linguistic Influences

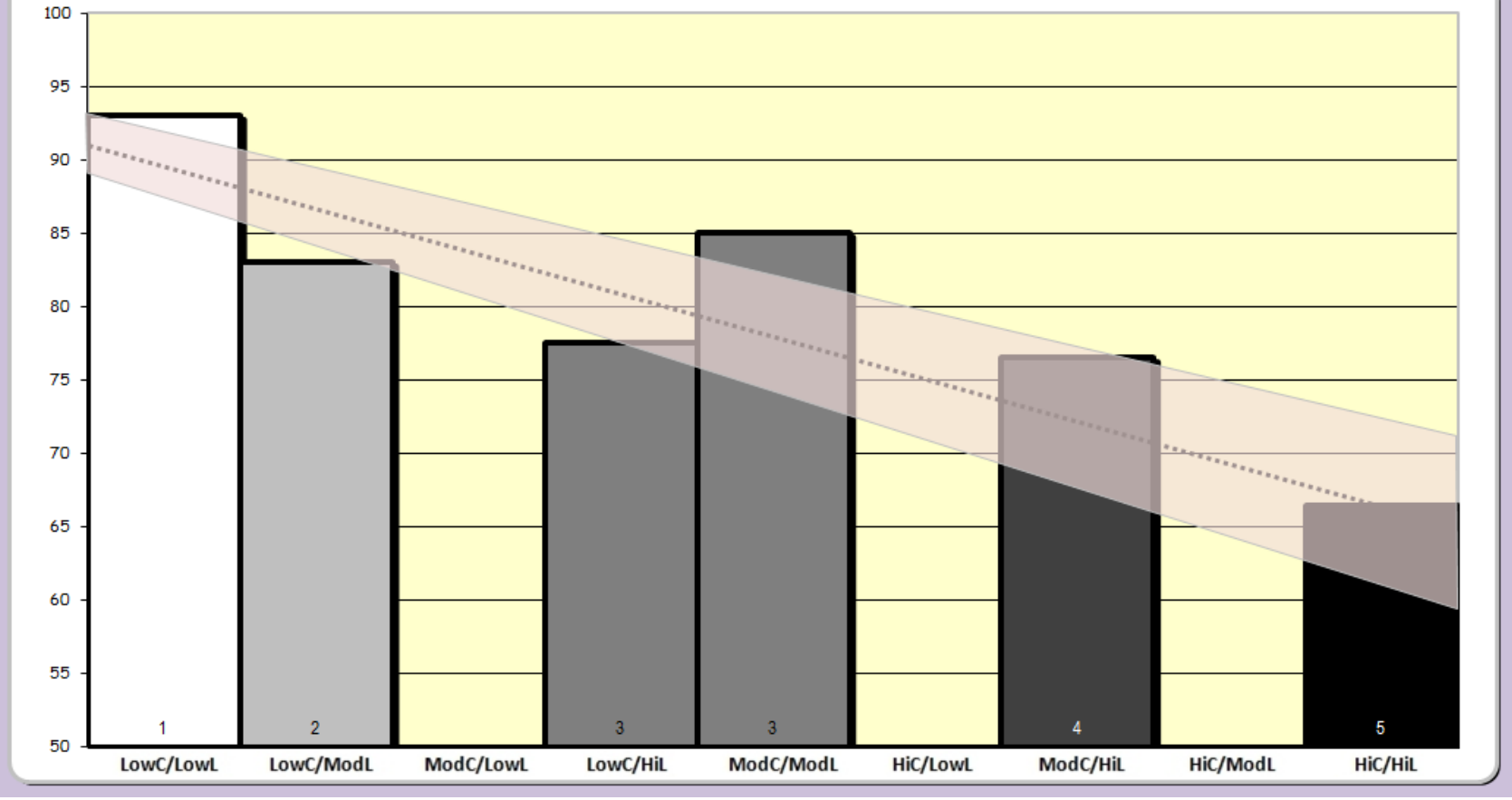


Culture-Language Interpretive Matrix: The Importance of the Context of Difference

Name: _____ Age: _____ Grade: _____

DIFFERENCE LEVEL FOR EVALUATION: ○ Slightly Different ○ Moderately Different ◉ Markedly Different

XBA C-LIM Graph for WJ III NU COG: Primary Evaluation of Cultural and Linguistic Influences



Evidence-based Assessment

Used in conjunction with other information relevant to appropriate bilingual, cross-cultural, nondiscriminatory assessment including…

- level of acculturation
- language proficiency
- socio-economic status
- academic history
- familial history
- developmental data
- work samples
- curriculum based data
- intervention results, etc.

…the C-LIM represents the only current method for evaluating test score validity and decreasing the potential for biased and discriminatory interpretation. When integrated within any assessment, the C-LIM facilitates evidence-based assessment and along with other assessment data, assists practitioners in answering the most basic question in EL assessment:

"Are the student's observed learning problems due primarily to cultural or linguistic differences or disorder?"

Assessment and Related Resources

TESTS:

Ortiz Picture Vocabulary Acquisition Test (Ortiz PVAT)
https://www.mhs.com/ortizpvat

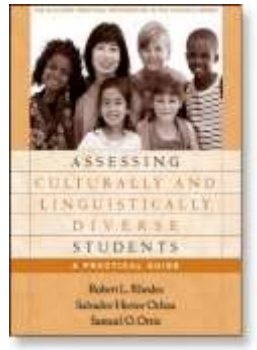
BOOKS:

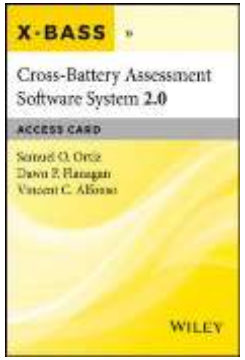
Rhodes, R., Ochoa, S. H. & Ortiz, S. O. (2005). Comprehensive Assessment of Culturally and Linguistically Diverse Students: A practical approach. New York: Guilford.

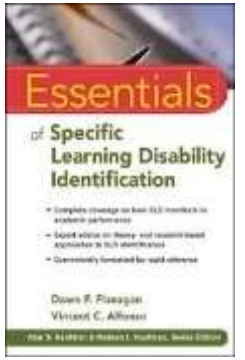
Flanagan, D. P., Ortiz, S.O. & Alfonso, V.C. (2013). Essentials of Cross-Battery Assessment, Third Edition. New York: Wiley & Sons, Inc.

Flanagan, D.P. & Ortiz, S.O. (2012). Essentials of Specific Learning Disability Identification. New York: Wiley & Sons, Inc.

Ortiz, S. O., Flanagan, D. P. & Alfonso, V. C. (2015). Cross-Battery Assessment Software System (X-BASS v1.4). New York: Wiley & Sons, Inc.







ONLINE:

Competency-based XBA Certification Program
https://www.schoolneuropsych.com/xba/

CHC Cross-Battery Online
http://www.crossbattery.com/

Free C-LIM Resources
http://facpub.stjohns.edu/~ortizs/CLIM/index.html

Cross-Battery Assessment XBA

ORTIZ PVAT™

ORTIZ PICTURE VOCABULARY ACQUISITION TEST™

Contemporary Measurement of Vocabulary Acquisition in an Era of Diversity:

Clinical and Educational Applications of the *Ortiz PVAT™*

For a free 30-day trial and two free uses:

http://info.mhs.com/OrtizPVATfreetrial

Free webinar now available at:

https://www.youtube.com/watch?v=EjUj0j_NIrQ

For additional information, visit:

http://www.mhs.com/ortizpvat

MHS® ASSESSMENTS