Assessment of English Learners:
Evidence-based evaluation and best practice.

Texas Association of School Psychologists

Houston, TX
October 25, 2019

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

TASP A partner on the path to student success
Texas Association of School Psychologists

1

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

The question regarding "difference vs. disorder" centers on the concept of **validity**.

2

Main Threats to Test Score Validity for ELLs

Acculturative Knowledge Acquisition – Not Item Content

"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 Race or Ethnicity

"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

3

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

Example of Potential Construct Invalidity:

"Assemble these blocks together in the correct manner so they appear identical to this illustration."

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.

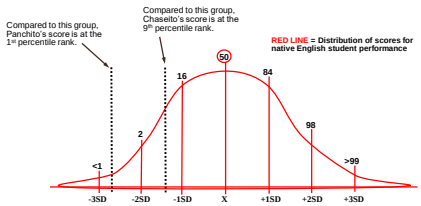
Example of Potential Interpretive Invalidity:

"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.

4

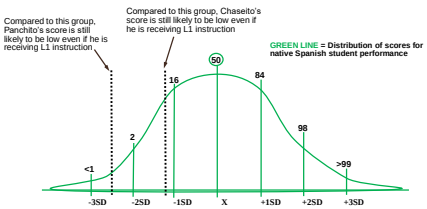
Diagnostic Question: Does Chaseito's or Panchito's rate of progress suggest cultural/linguistic difference or possible disorder?



For the purposes of determining whether a disability exists, use of a monolingual English speaking comparison group is discriminatory and makes it appear incorrectly that both students might have some type of disability.

5

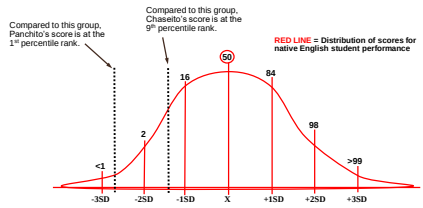
Diagnostic Question: Does Chaseito's or Panchito's rate of progress suggest cultural/linguistic difference or possible disorder?



Similarly, use of a monolingual, native-language speaking group remains discriminatory because neither student is monolingual anymore (even when receiving native language instruction) and it continues to make it appear incorrectly that both Chaseito and Panchito have some type of disability.

6

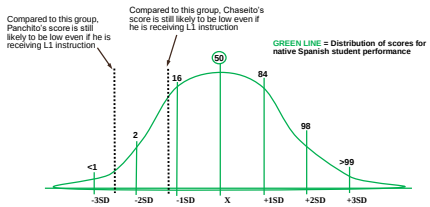
Intervention Question: What are Chaseito's and Panchito's instructional levels, needs, goals, and how far behind are they academically?



This is a valid comparison, but only to the extent that it shows that both students are well below grade-level academic standards. However, this comparison only informs their instructional level and it cannot be used to identify any of the possible reasons regarding they are low, only that they are.

7

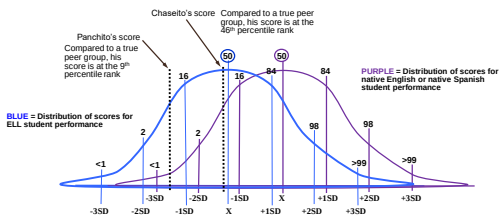
Intervention Question: What are Chaseito's and Panchito's instructional levels, needs, goals, and how far behind are they academically?



Use of a native-language comparison standard remains inappropriate, even in cases where native language (L1) instruction is provided because they are no longer monolingual. Simply entering the U.S. school system and being expected to learn English invalidates use of monolingual L1 samples and continues to be discriminatory.

8

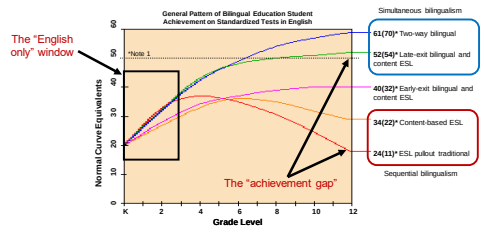
Diagnostic Question: Does Chaseito's or Panchito's rate of progress suggest cultural/linguistic difference or possible disorder?



Whether conducted through RTI/MTSS or testing, only use of a "true peer" comparison group provides the basis for making non-discriminatory diagnostic decisions as long as there is control for developmental language differences between English learners and English speakers and among English learners and other English learners.

9

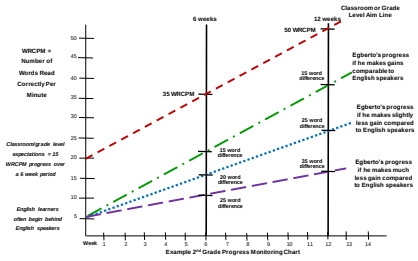
Academic Test Score Validity Requires "True Peer" Comparison



*Note 1: Average performance of native-English speakers making one year's progress in each grade. Scores in parentheses are percentile ranks converted from NCEs. Adapted from: Thomas, W. & Collier, V. (1997). Language Minority Student Achievement and Program Effectiveness. Washington DC: National Clearinghouse for Bilingual Education.

10

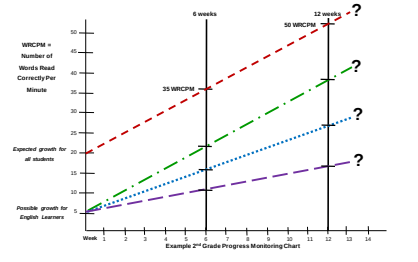
Fairness in Evaluation of ELs via RTI/MTSS



*Note. The name, "Egberto," is a derivative of "Egbert" and used with the blessings of Dan Reschly.

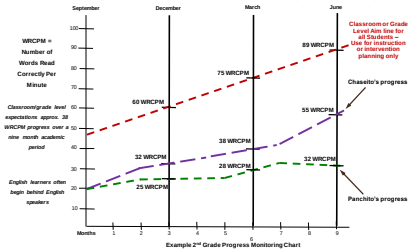
11

The Most Appropriate Standard for Comparison Depends on the Question Being Asked: An RTI example.



12

Intervention Question: What are Chaseito's and Panchito's instructional levels, needs, goals, and how far behind are they academically?

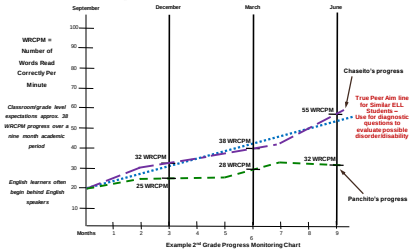


ANSWER: Both Chaseito and Panchito are significantly behind grade level expectations. Both need systematic, high-quality instruction consistent with their language proficiency to promote continued growth in reading toward grade level standards.

Thus, comparison to native English speakers is appropriate only for questions related to instructional need, intervention planning, and programming goals, but is NOT appropriate for questions about possible disability where it would be discriminatory.

13

Diagnostic Question: Does Chaseito's or Panchito's rate of progress suggest cultural/linguistic difference or possible disorder?



ANSWER: Chaseito's rate of progress and development is commensurate with that of similar, same age peers and does not suggest any problems, however, Panchito's rate of progress is below that expected of same age peers and may suggest a disorder.

Thus, to avoid being discriminatory, comparison to other TRUE PEER English learners is necessary for any diagnostic questions related to possible disorder or disability. It may also add information related to instructional needs and intervention.

14

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 relative 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 & Frey, 2012

15

Processes and Procedures for Addressing Test Score Validity

In what manner exactly, is evidence-based, nondiscriminatory assessment conducted and to what extent is there any research to support the use of any of the following methods as being capable of establishing sufficient test score validity?

- **Modified Methods of Evaluation**
  - *Working around the language by modifying/altering the assessment*
- **Nonverbal Methods of Evaluation**
  - *Avoiding the language by evaluating areas unrelated to language*
- **Dominant Language Evaluation**
  - *Choosing a language based simply on relative proficiency*

16

Processes and Procedures for Addressing Test Score Validity

ISSUES IN MODIFIED METHODS OF EVALUATION

Modified and Altered Assessment: Just workaround the language.

- *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***

17

Processes and Procedures for Addressing Test Score Validity

ISSUES IN NONVERBAL METHODS OF EVALUATION

Language Reduced Assessment: Just avoid the language.

- *"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***

18

Processes and Procedures for Addressing Test Score Validity

ISSUES IN DOMINANT LANGUAGE EVALUATION

Determining the language of evaluation: Just choose a language.

- *generally refers to the assessment of an EL after it has been determined that the examinee is more proficient ("dominant") in one language than the other*
- *being "dominant" in a language does not imply age-appropriate development in that language*
- *dominance does not inform instructional intervention, progress, growth, or expected test performance*
- *dominance is often affected by preferences that are shaped by social factors including identity development*
- *direct evaluation in the native language (L1) can only be conducted by a bilingual evaluator and is not an option available to monolingual English speaking evaluators*
- *bilingual ability is no guarantee of nondiscriminatory assessment--native language assessment (L1) can be just as biased and inequitable as assessment in English (L2)*
- *in contrast to assessment in English, native language evaluation assessment is a relatively new idea without a substantive empirical base to guide or support standards of practice*
- *both L1 and L2 test norm samples fail to control for variability between and among ELs relative to their own amount of exposure to English and to that of monolingual, native English speaker*
- ***without a research base, there is no way to evaluate the validity of test results derived simply by testing in the dominant language** and any subsequent interpretations would be specious and amount to no more than a guess*

19

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.

20

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

Whatever method or approach may be employed in evaluation of EL's, the fundamental obstacle to nondiscriminatory interpretation rests on the degree to which the examiner is able to defend claims of test score (construct and interpretive) 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.

Test score validity must be evaluated or established via use of a "true peer" comparison standard and the only manner in which to accomplish this task is with evidence and data.

21

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.

22

Summary of Research on the Test Performance of English Language Learners

Research conducted over the past 100 years on ELs who are non-disabled, of average ability, possess moderate to high proficiency in English, and tested in English, has resulted in a basic and ubiquitous finding:

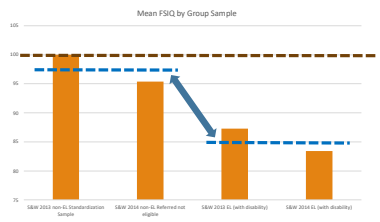
English Learners and Native English speakers tend to perform differently on standardized, norm-referenced tests of intelligence and general cognitive ability.

So what explains these findings? Early explanations relied on genetic differences attributed to racial inferiority. But even early researchers noticed that language differences (i.e., lack of proficiency) likely played a role in this difference, particularly because ELs also tended to perform better on nonverbal tests than on verbal tests.

23

Research Foundations for EL Evaluation

ELs and non-EL's perform differently: Broad ability level

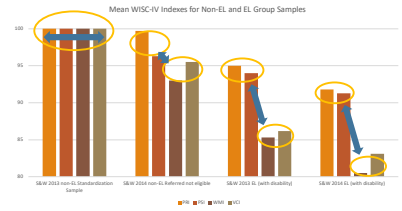


Styck, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382.

24

Research Foundations for EL Evaluation

ELs and non-EL's perform differently: Index level

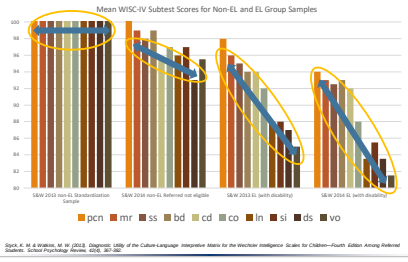


Stock, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382.

25

Research Foundations for EL Evaluation

ELs and non-EL's perform differently: Subtest level



Stock, 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.

26

Research Foundations for EL Evaluation

Although it has long been recognized that ***language*** likely account for the differences in test performance between English leaners and native English speakers, its influence has rarely been examined directly as a confounding variable and there has been a tendency instead to use "cultural" and "racial/ethnic" variables as proxies for language.

EL vs. ES: In general, research with ELs indicates that language (including acquisition of acculturative knowledge) has a powerful and significant effect on test performance that can be discerned at every level of testing, broad ability, index/composite, or subtest.

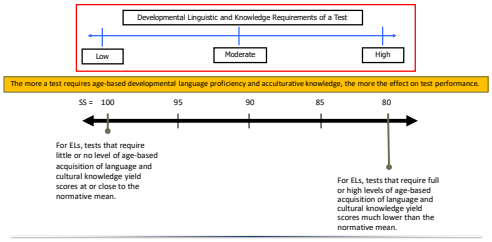
EL vs. EL: In addition, differences in exposure to and development in English varies among ELs such that the influence increases proportionally on tests that use, measure, and rely more on language and language-based abilities.

When understood as such, the impact of language on test performance of ELs is not seen to be a simple "verbal vs. nonverbal" dichotomy but rather ***a continuum formed by a linear and proportional attenuation of performance relative to both ESs and other ELs.***

27

Research Foundations for EL Evaluation: EL to ES

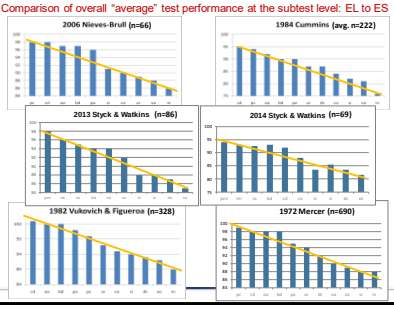
When compared to monolingual, native-English speakers, language influences EL test performance in a linear, continuous manner, not dichotomously, across all subtests.



28

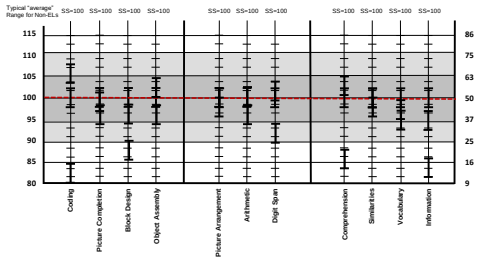
Research Foundations for EL Evaluation: EL to ES

Comparison of overall "average" test performance at the subtest level: EL to ES



29

Fairness in Determining "Average" Performance: ES to ES



30

Research Foundations for EL Evaluation: EL to ES

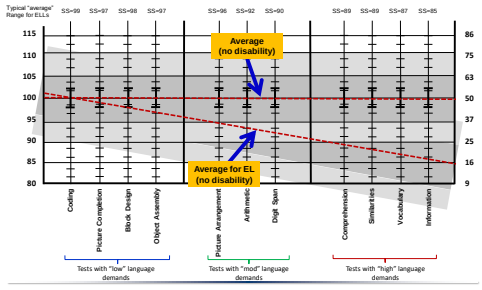
EL performance is moderated by level of English proficiency as compared to ES

	Subtest Name	Mercer 1972 Scale Score	Vukovich & Figueroa, 1982 Scale Score	Cummins 1982 Scale Score	Nieves-Brull 2006 Scale Score	Grand Mean
Tests with **"high"** language demands	Information	7.5	7.8	5.1	7.2	85
	Vocabulary	8.0	8.3	6.1	7.5	87
	Similarities	7.6	8.8	6.4	8.2	89
	Comprehension	7.8	9.0	6.7	8.0	89
Tests with **"mod"** language demands	Digit Span	8.3	8.5	7.3	*	90
	Arithmetic	8.7	9.4	7.4	7.8	92
	Picture Arrangement	9.0	10.3	8.0	9.2	96
Tests with **"low"** language demands	Block Design	9.5	10.8	8.0	9.4	97
	Object Assembly	9.6	10.7	8.4	9.3	98
	Picture Completion	9.7	9.9	8.7	9.5	97
	Coding	9.6	10.9	8.9	9.6	99

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

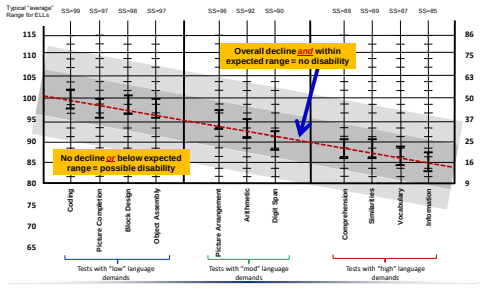
31

Fairness in Determining "Average" Performance: EL to ES



32

Fairness in Determining "Average" Performance: EL to ES



33

Interpretive Errors in C-LIM Studies: Styck & Watkins

Overall decline *and* within expected range = no disability

No decline *or* below expected range = possible disability

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

The authors noted that *"roughly 97% of (n = 83) of participants were identified as meeting criteria for an educational disability (86% as SLD)" (p. 371).* Yet, only 9 ELL cases (10.5%) resulted in invalid scores (no disability). Thus, the C-LIM suggested invalid scores in 9 cases, 3 of which were likely correct (those without disabilities) so that the C-LIM was consistent with and supported the placement decision of the child by the district in 93% of the cases (89.5% + 3.5%). Moreover, the results of analyses with the WISC-IV normative sample show that declines relative to language are unusual, perhaps even indications of potential SLI in monolingual, native English speakers as described by Cormier et al. (2014).

To summarize, far from undermining the validity of the C-LIM, the Styck & Watkins studies provide strong and powerful support for the clinical utility and validity of the C-LIM when evaluating EL test performance.

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

34

Research Foundations for EL Evaluation: EL to ES

The influence of language on subtest level performance in English speakers and English learners.

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

	Individual test	Variance explained 7-10	11-14	15-18	
Highest Language Demands	Verbal Comprehension	.79[a]	.86[a]	.81[a]	C-LIM Level 5
	General Information	.71[a]	.85[a]	.86[a]	
	Concept Formation	.67[a]	.71[a]	.67[a]	
	Visual–Auditory Learning	.40[b]	.37[b]	.41[b]	C-LIM Level 4
	Delayed Recall Visual–Auditory Learning	.39[b]	.32[b]	.37[b]	
	Analysis Synthesis	.29[b]	.44[b]	.47[b]	
	Sound Blending	.25[b]	.32[b]	.35[b]	
	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]	C-LIM Level 3
	Numbers Reversed	.17[b]	.26[b]	.30[b]	
	Pair Cancelation	.17[b]	.11[b]	.11[b]	
	Rapid Picture Naming	.16[b]	.07[b]	.16[b]	
	Incomplete Words	.13[b]	.31[b]	.23[b]	
	Visual Matching	.13[b]	.15[b]	.16[b]	C-LIM Level 2
	Decision Speed	.12[b]	.15[b]	.19[b]	
	Auditory Attention	.10[b]	.20[b]	.15[b]	
	Spatial Relations	.09[b]	.16[b]	.16[b]	
Lowest Language Demands	Planning	.07[b]	.12[b]	.11[b]	C-LIM Level 1
	Picture Recall	.02[b]	.06[b]	.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.*

35

Research Foundations for EL Evaluation: EL to ES

EL performance is moderated by level of English proficiency as compared to ES

	Subtest Name	Mercer 1972 Scale Score	Vukovich & Figueroa, 1982 Scale Score	Cummins 1982 Scale Score	Nieves-Brull 2006 Scale Score	Grand Mean	C-LIM Level
Tests with "high" language demands	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
Tests with "mod" language demands	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
Tests with "low" language demands	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.*

36

Fairness in Determining "Average" Performance: EL to ES

Matrix of WISC subtest means arranged by EL vs. ES test performance

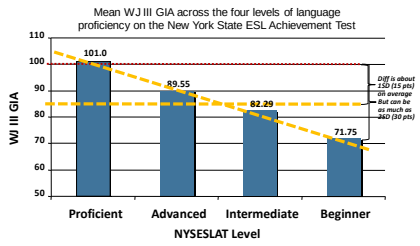
DEGREE OF LINGUISTIC DEMAND

DEGREE OF CULTURAL LOADING	LOW	MODERATE	HIGH
LOW	Coding Object Assembly Level 1 SS= 99	Block Design Level 2 SS= 97	Digit Span Level 3 SS= 91
MODERATE	Picture Completion Level 2 SS= 97	Arithmetic Level 3 SS= 91	Comprehension Level 4 SS= 89
HIGH	Picture Arrangement Level 3 SS= 91	Level 4 SS= 89	Information Similarities Vocabulary Level 5 SS= 85

37

Research Foundations for EL Evaluation: EL to EL

General ability level performance as compared to other English learners

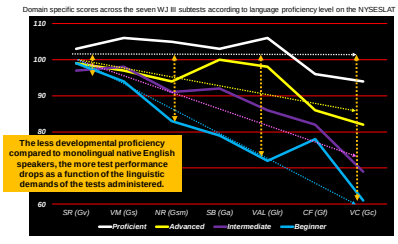


Source: Sotelo-Dynega, M., Ortiz, S.O., Flanagan, D.P., Chaplin, W. (2013).

38

Research Foundations for EL Evaluation: EL to EL

Subtest level performance as compared to other English Learners



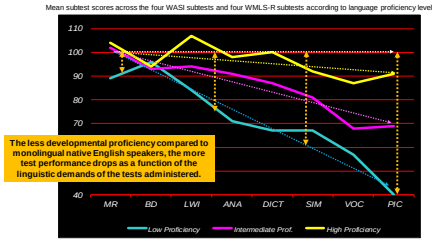
Source: Sotelo-Dynega, M., Ortiz, S.O., Flanagan, D.P., Chaplin, W. (2013). English Language Proficiency and Test Performance: Evaluation of bilinguals with the Woodcock-Johnson III Tests of Cognitive Ability. Psychology in the Schools, Vol 50(8), pp. 781-797.

39

Research Foundations for EL Evaluation: EL to EL

Subtest level performance as compared to other English Learners

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.

40

Summary of Research Foundations for EL Evaluation

1. COMPARED TO ENGLISH SPEAKERS (EL to ES): *Test performance of ELs is moderated by the degree to which a given index or subtest relies on or requires age- or grade-expected English language development and the acquisition of incidental acculturative knowledge.*

2. COMPARED TO ENGLISH LEARNERS (EL to EL): *Test performance of ELs is further moderated by the degree to which an EL varies in terms of their own developmental English language proficiency and acculturative knowledge acquisition.*

Proper interpretation of EL test performance thus requires a true peer group of other ELs that is based not on the language spoken by the individual but on comparison to other ELs with the same degree of English exposure and development.

With one exception, current test norm samples lack control for developmental differences in English language exposure. This means that interpretation of test scores at any level must be made within the context of research which provides the only empirically-derived, albeit, very rough, true peer standard or "norm group".

Use of research on the relative test performance of ELs based on language exposure (as reflected by the degree of "difference" the student displays relative to the norm samples of the tests being used) is the very foundation and sole purpose of the C-LIM.

41

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.

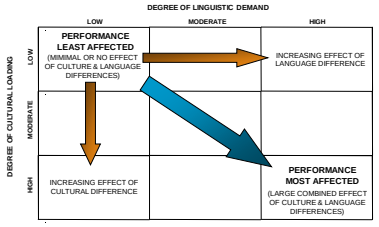
The primary purpose of the C-LIM is to assist evaluators in ruling out cultural and linguistic influences as exclusionary factors that may have undermined the validity of test scores, particularly in evaluations of SLD or other cognitive-based disorders. Being able to make this determination is the primary 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.

Free version of C-LIM available at: http://facpub.stjohns.edu/~ortizs/CLIM/

42

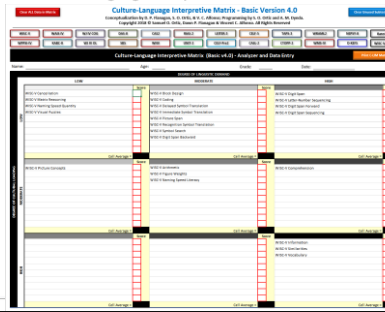
Fairness in Determining "Average" Performance: EL to ES

Matrix arrangement of expected subtest level performance for ELs vs. ES



43

The Culture-Language Interpretive Matrix – Basic Version 4.0



44

Fairness in Determining "Average" Performance: EL to EL

Research-based subtest means regarding expected test performance EL vs. EL

Degree of Cultural Loading	Degree of Linguistic Demand: 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

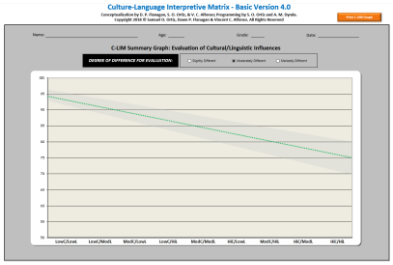
Slightly Different: *Includes individuals with very high levels of English language proficiency (e.g., CALP) and high acculturation, but still in ot entirely comparable to mainstream U.S. English speakers. Examples include individuals who are third generation in the U.S., have well educated/higher SES 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.*

45

The Culture-Language Interpretive Matrix – Basic Version 4.0



46

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.

47

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.

Results are **INVALID** only if both conditions are met.

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. If either criterion is not met, the results can be assumed to be **VALID**.

48

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?

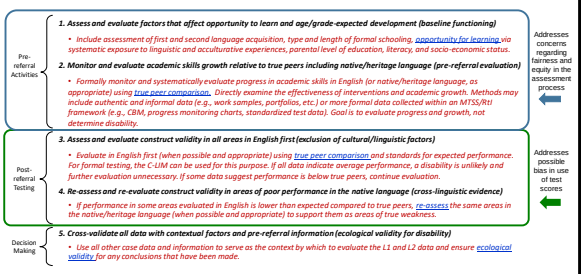
3. Is the pattern of scores consistent with a developmental explanation of the examinee's educational program and experiences?

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.

49

A Best Practice Framework for Comprehensive Evaluation of ELs



50

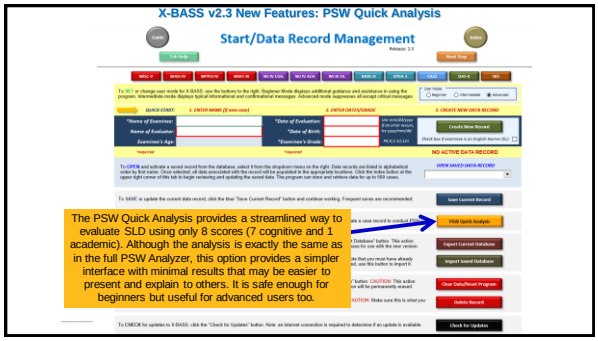
Practical Considerations for Addressing Test Score Validity in Evaluation of ELs

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*

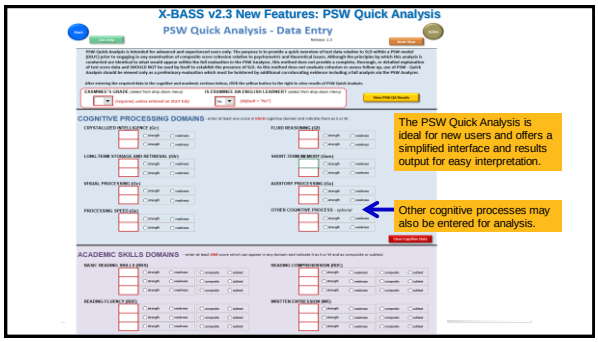
51


Translating Research into Practice
Evaluation Issues and Methods
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 EL test performance, they can be used to define and establish a "true peer" reference group for disability-based evaluations.

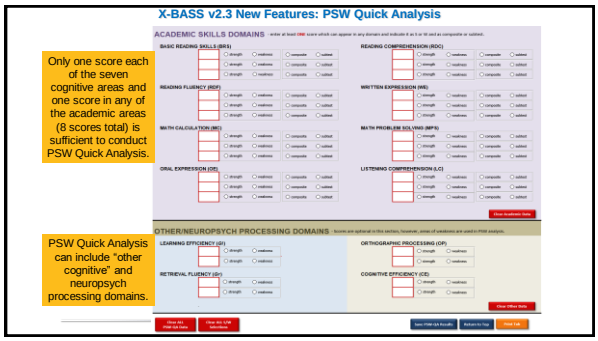
52


X-BASS v2.3 New Features: PSW Quick Analysis
Start/Data Record Management
The PSW Quick Analysis provides a streamlined way to evaluate SLD using only 8 scores (7 cognitive and 1 academic). Although the analysis is exactly the same as in the full PSW Analyzer, this option provides a simpler interface with minimal results that may be easier to present and explain to others. It is safe enough for beginners but useful for advanced users too.

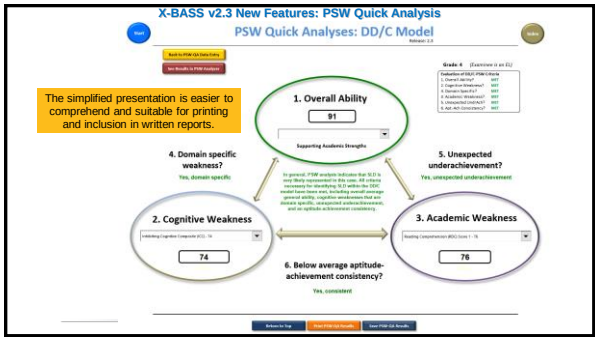
53


X-BASS v2.3 New Features: PSW Quick Analysis
PSW Quick Analysis - Data Entry
The PSW Quick Analysis is ideal for new users and offers a simplified interface and results output for easy interpretation.
Other cognitive processes may also be entered for analysis.

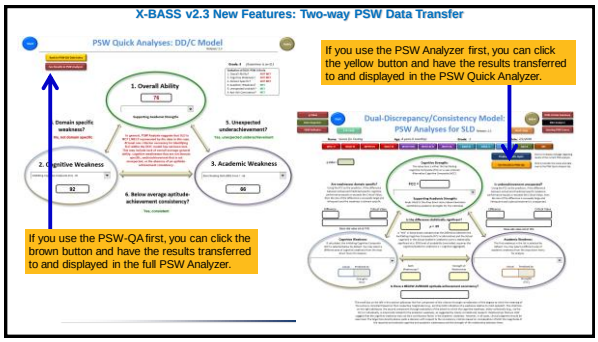
54


X-BASS v2.3 New Features: PSW Quick Analysis
Only one score each of the seven cognitive areas and one score in any of the academic areas (8 scores total) is sufficient to conduct PSW Quick Analysis.
PSW Quick Analysis can include "other cognitive" and neuropsych processing domains.
ACADEMIC SKILLS DOMAINS
OTHER/NEUROPSYCH PROCESSING DOMAINS

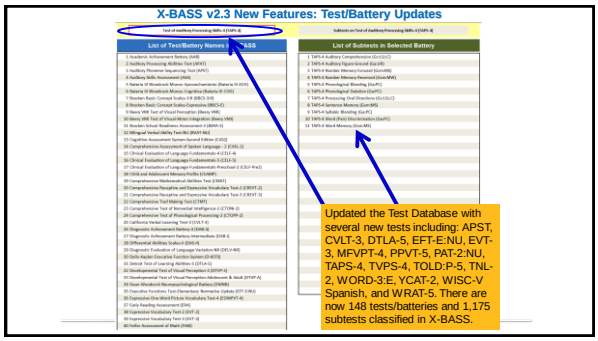
55


X-BASS v2.3 New Features: PSW Quick Analysis
PSW Quick Analyses: DD/C Model
The simplified presentation is easier to comprehend and suitable for printing and inclusion in written reports.
1. Overall Ability
91
4. Domain specific weakness?
Yes, domain specific
5. Unexpected underachievement?
Yes, unexpected underachievement
2. Cognitive Weakness
74
3. Academic Weakness
76
6. Below average aptitude-achievement consistency?
Yes, consistent

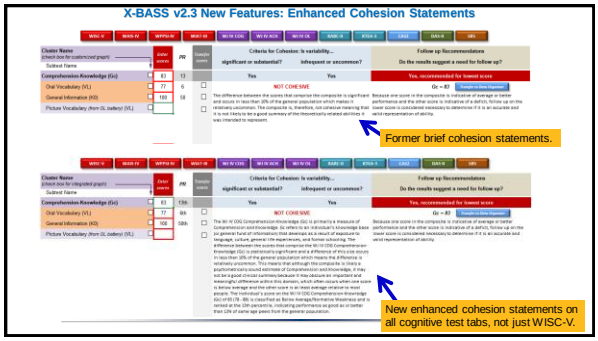
56


X-BASS v2.3 New Features: Two-way PSW Data Transfer
PSW Quick Analyses: DD/C Model
If you use the PSW Analyzer first, you can click the yellow button and have the results transferred to and displayed in the PSW Quick Analyzer.
If you use the PSW-QA first, you can click the brown button and have the results transferred to and displayed in the full PSW Analyzer.
Dual-Discrepancy/Consistency Model: PSW Analyses for SLD

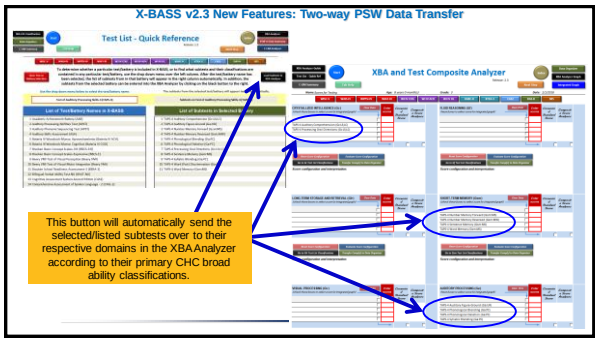
57


X-BASS v2.3 New Features: Test/Battery Updates
List of Tests/Battery Names in X-BASS
List of Subtests in Selected Battery
Updated the Test Database with several new tests including: APST, CVLT-3, DTLA-5, EFT-E:NU, EVT-3, MFVPT-4, PPVT-5, PAT-2:NU, TAPS-4, TVPS-4, TOLD:P-5, TNL-2, WORD-3:E, YCAT-2, WISC-V Spanish, and WRAT-5. There are now 148 tests/batteries and 1,175 subtests classified in X-BASS.

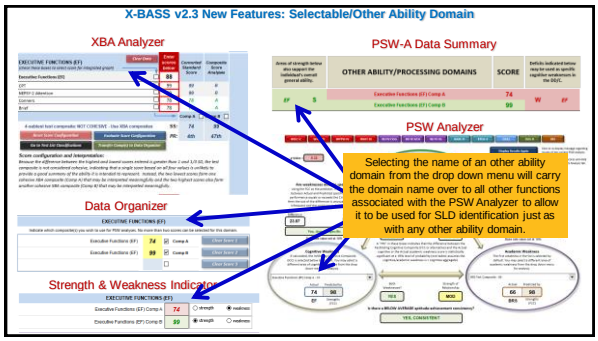
58


X-BASS v2.3 New Features: Enhanced Cohesion Statements
Former brief cohesion statements.
New enhanced cohesion statements on all cognitive test tabs, not just WISC-V.

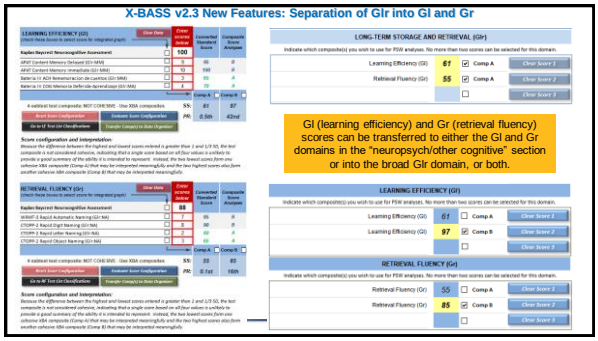
59


X-BASS v2.3 New Features: Two-way PSW Data Transfer
Test List - Quick Reference
XBA and Test Composite Analyzer
This button will automatically send the selected/listed subtests over to their respective domains in the XBA Analyzer according to their primary CHC broad ability classifications.

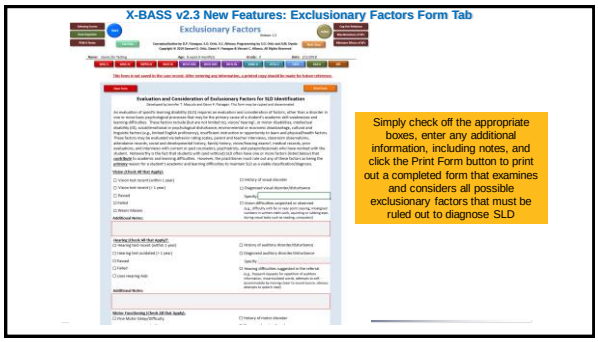
60


X-BASS v2.3 New Features: Selectable/Other Ability Domain
XBA Analyzer
PSW-A Data Summary
PSW Analyzer
Data Organizer
Strength & Weakness Indicator
Selecting the name of an other ability domain from the drop down menu will carry the domain name over to all other functions associated with the PSW Analyzer to allow it to be used for SLD identification just as with any other ability domain.

61


X-BASS v2.3 New Features: Separation of Glr into Gl and Gr
GI (learning efficiency) and Gr (retrieval fluency) scores can be transferred to either the Gl and Gr domains in the "neuropsych/other cognitive" section or into the broad Glr domain, or both.

62


X-BASS v2.3 New Features: Exclusionary Factors Form Tab
Exclusionary Factors
Simply check off the appropriate boxes, enter any additional information, including notes, and click the Print Form button to print out a completed form that examines and considers all possible exclusionary factors that must be ruled out to diagnose SLD

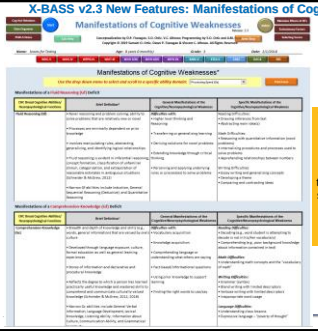
63

X-BASS v2.3 New Features: Cognitive-Achievement Relations Tab

This new tab contains a table that provides information regarding the relationship between an academic area (and subskill) to specific areas of cognitive functioning. An explanation of the possible etiology is also provided.

64

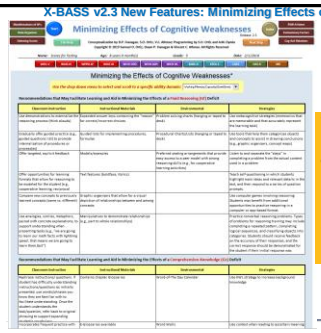
X-BASS v2.3 New Features: Manifestations of Cognitive Weaknesses Tab



This new tab contains a table that provides a definition of academically-related cognitive abilities as well as their general and specific manifestations in terms of academic functioning and skills acquisition. The table can be quickly navigated by selecting the cognitive domain from the drop down menu at the top.

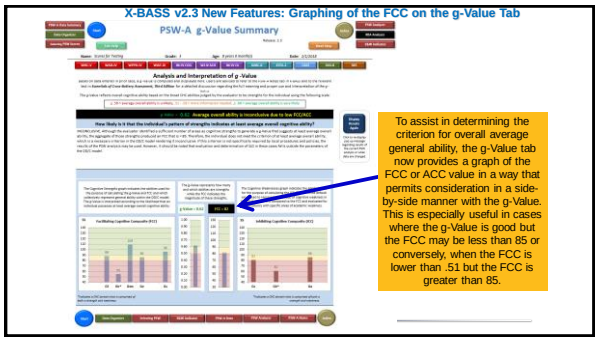
65

X-BASS v2.3 New Features: Minimizing Effects of Cognitive Weaknesses Tab

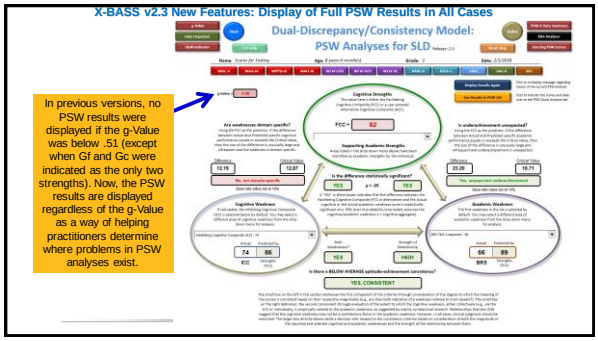


This new tab contains a table that provides information regarding instructional, environmental, and other strategies for minimizing the effects of cognitive weaknesses which may be helpful in determining appropriate avenues for intervention. The table can be quickly navigated by selecting the cognitive domain from the drop down menu at the top.

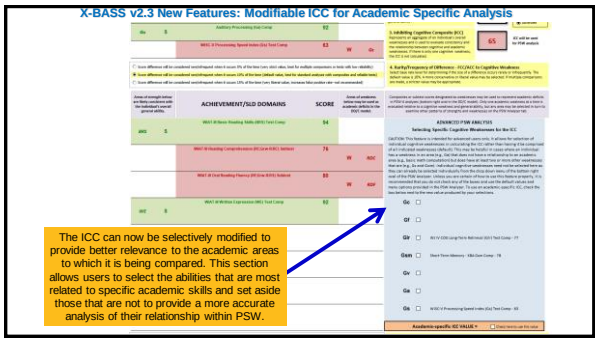
66


X-BASS v2.3 New Features: Graphing of the FCC on the g-Value Tab
PSW-A g-Value Summary
To assist in determining the criterion for overall average general ability, the g-Value tab now provides a graph of the FCC or ACC value in a way that permits consideration in a side-by-side manner with the g-Value. This is especially useful in cases where the g-Value is good but the FCC may be less than 85 or conversely, when the FCC is lower than .51 but the FCC is greater than 85.

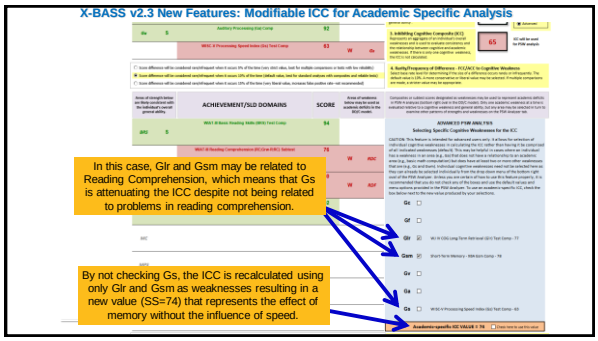
67


X-BASS v2.3 New Features: Display of Full PSW Results in All Cases
Dual-Discrepancy/Consistency Model:
PSW Analyses for SLD
In previous versions, no PSW results were displayed if the g-Value was below .51 (except when Gf and Gc were indicated as the only two strengths). Now, the PSW results are displayed regardless of the g-Value as a way of helping practitioners determine where problems in PSW analyses exist.

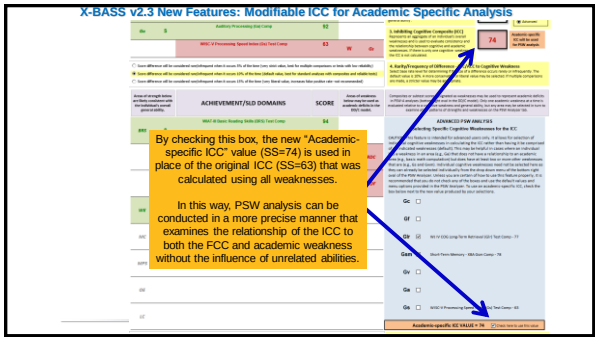
68


X-BASS v2.3 New Features: Modifiable ICC for Academic Specific Analysis
The ICC can now be selectively modified to provide better relevance to the academic areas to which it is being compared. This section allows users to select the abilities that are most related to specific academic skills and set aside those that are not to provide a more accurate analysis of their relationship within PSW.

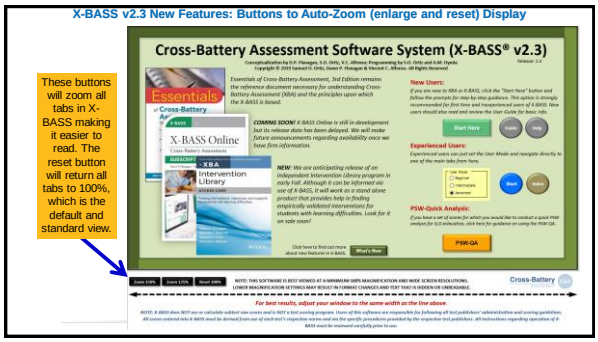
69


X-BASS v2.3 New Features: Modifiable ICC for Academic Specific Analysis
In this case, Glr and Gsm may be related to Reading Comprehension, which means that Gs is attenuating the ICC despite not being related to problems in reading comprehension.
By not checking Gs, the ICC is recalculated using only Glr and Gsm as weaknesses resulting in a new value (SS=74) that represents the effect of memory without the influence of speed.

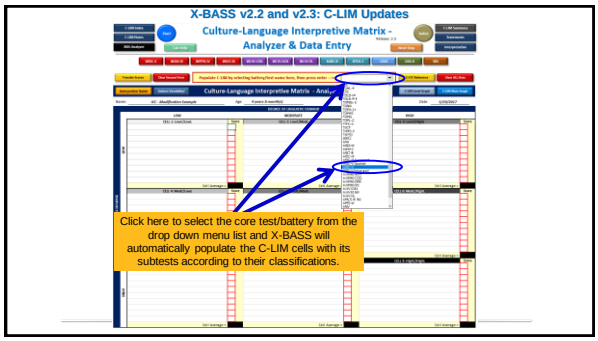
70


X-BASS v2.3 New Features: Modifiable ICC for Academic Specific Analysis
By checking this box, the new "Academic-specific ICC" value (SS=74) is used in place of the original ICC (SS=63) that was calculated using all weaknesses.
In this way, PSW analysis can be conducted in a more precise manner that examines the relationship of the ICC to both the FCC and academic weakness without the influence of unrelated abilities.

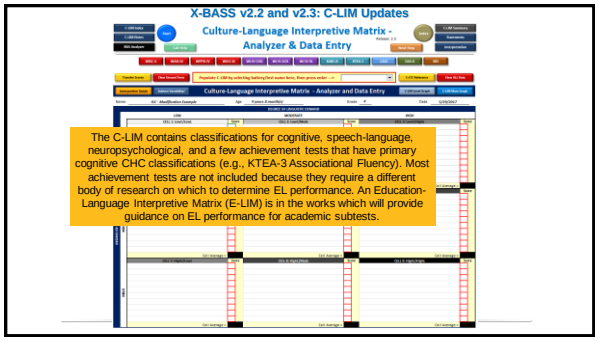
71


X-BASS v2.3 New Features: Buttons to Auto-Zoom (enlarge and reset) Display
These buttons will zoom all tabs in X-BASS making it easier to read. The reset button will return all tabs to 100%, which is the default and standard view.
Cross-Battery Assessment Software System (X-BASS® v2.3)

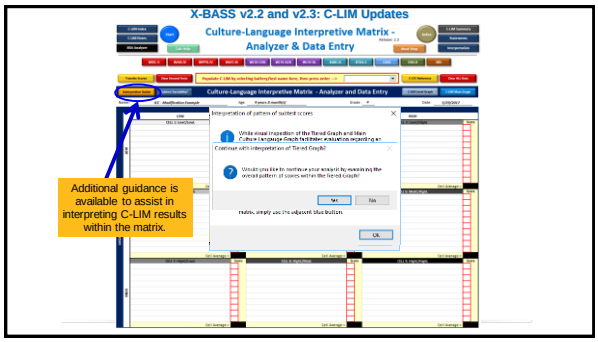
72


X-BASS v2.2 and v2.3: C-LIM Updates
Culture-Language Interpretive Matrix -
Analyzer & Data Entry
Click here to select the core test/battery from the drop down menu list and X-BASS will automatically populate the C-LIM cells with its subtests according to their classifications.

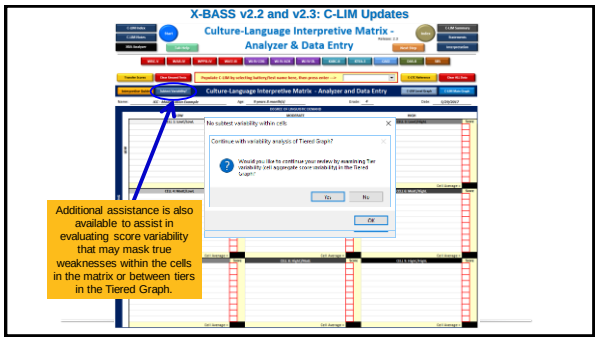
73


X-BASS v2.2 and v2.3: C-LIM Updates
Culture-Language Interpretive Matrix -
Analyzer & Data Entry
The C-LIM contains classifications for cognitive, speech-language, neuropsychological, and a few achievement tests that have primary cognitive CHC classifications (e.g., KTEA-3 Associational Fluency). Most achievement tests are not included because they require a different body of research on which to determine EL performance. An Education-Language Interpretive Matrix (E-LIM) is in the works which will provide guidance on EL performance for academic subtests.

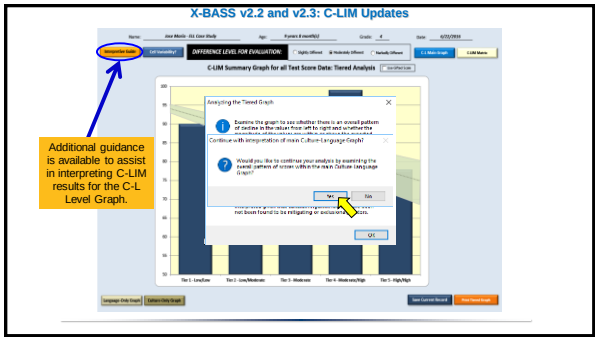
74


X-BASS v2.2 and v2.3: C-LIM Updates
Culture-Language Interpretive Matrix -
Analyzer & Data Entry
Additional guidance is available to assist in interpreting C-LIM results within the matrix.

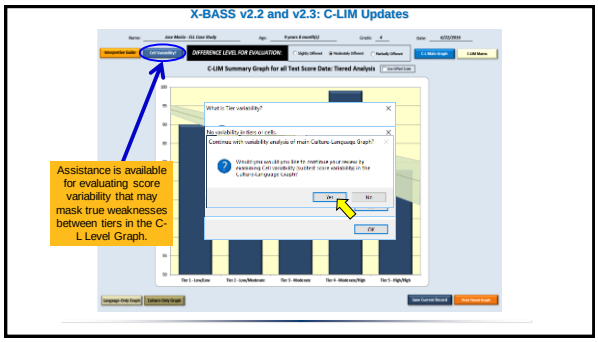
75


X-BASS v2.2 and v2.3: C-LIM Updates
Culture-Language Interpretive Matrix - Analyzer & Data Entry
Additional assistance is also available to assist in evaluating score variability that may mask true weaknesses within the cells in the matrix or between tiers in the Tiered Graph.

76

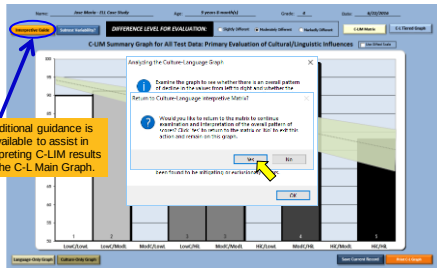

X-BASS v2.2 and v2.3: C-LIM Updates
Additional guidance is available to assist in interpreting C-LIM results for the C-L Level Graph.

77


X-BASS v2.2 and v2.3: C-LIM Updates
Assistance is available for evaluating score variability that may mask true weaknesses between tiers in the C-L Level Graph.

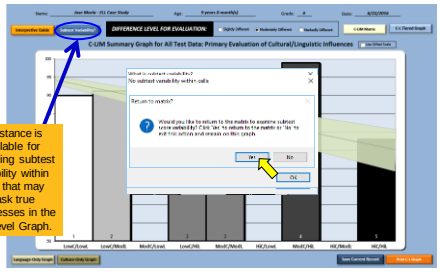
78

X-BASS v2.2 and v2.3: C-LIM Updates



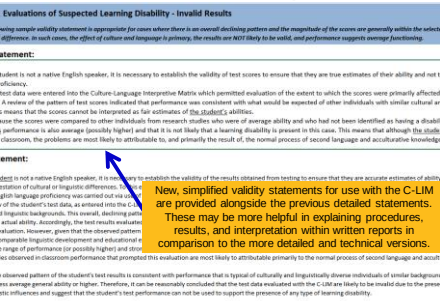
79

X-BASS v2.2 and v2.3: C-LIM Updates



80

X-BASS v2.2 and v2.3: C-LIM Updates



81

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

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

82

Multilingual Assessment of ELs: Step by Step

Step 1. Test first in English (L2) and evaluate construct validity in all areas in English (exclusion of cultural/linguistic factors)

- *If all scores indicate normative strengths (SS ≈ 90 or higher) when tested in English (L2), scores are valid to the extent that a disability is not likely, thus no further testing is necessary.*
- *If some scores are normative weaknesses (SS < ≈ 90) evaluate test score validity in a research-based manner, e.g., via the C-LIM.*
- *If C-LIM indicates primary influence of language/culture, test scores are likely invalid and indicate average ability in all areas and a disability is not likely, thus no further testing is necessary.*
- *If C-LIM indicates contributory or minimal influence of language/culture, test scores are likely to be valid and the evaluation should continue.*

Step 2. Re-evaluate areas of weakness in native language (L2) to provide additional supporting evidence of validity (cross-linguistic confirmation)

- *If data indicate an area is a strength (i.e., average), then original L2 score is invalid, use the L1 score.*
- *If data indicate an area is still a weakness, then original L2 score is valid, use the L2 score.*

Step 3. Further cross-validate L1 and L2 test scores with contextual factors and pre-referral data and academic concerns (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*

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. However, the approach does require knowledge of research on ELs and a systematic way to evaluate test score validity via true peer group comparison.

83

Multilingual Assessment of ELs: Step by Step

Step 1. Test first in English (L2) and evaluate construct validity in all areas in English (exclusion of cultural/linguistic factors)

- *If all scores indicate normative strengths (SS ≈ 90 or higher) when tested in English (L2), scores are valid to the extent that a disability is not likely, thus no further testing is necessary.*
- *If some scores are normative weaknesses (SS < ≈ 90) evaluate test score validity in a research-based manner, e.g., via the C-LIM.*
- *If C-LIM indicates primary influence of language/culture, test scores are likely invalid and indicate average ability in all areas and a disability is not likely, thus no further testing is necessary.*
- *If C-LIM indicates contributory or minimal influence of language/culture, test scores are likely to be valid and the evaluation should continue.*

Step 2. Re-evaluate areas of weakness in native language (L2) to provide additional supporting evidence of validity (cross-linguistic confirmation)

- *If data indicate an area is a strength (i.e., average), then original L2 score is invalid, use the L1 score.*
- *If data indicate an area is still a weakness, then original L2 score is valid, use the L2 score.*

Step 3. Further cross-validate L1 and L2 test scores with contextual factors and pre-referral data and academic concerns (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*

84

Basic Disability Evaluation with an English Learner: A Case Study

WISC-V/WJ IV/WIAT-III XBA DATA FOR Maria Ayala
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

85

Multilingual Assessment of ELs: Step by Step

Step 1. Test first in English (L2) and evaluate construct validity in all areas in English (exclusion of cultural/linguistic factors)

- *If all scores indicate normative strengths (SS ≈ 90 or higher) when tested in English (L2), scores are valid to the extent that a disability is not likely, thus no further testing is necessary.*
- *If some scores are normative weaknesses (SS < ≈ 90) evaluate test score validity in a research-based manner, e.g., via the C-LIM.*
- *If C-LIM indicates primary influence of language/culture, test scores are likely invalid and indicate average ability in all areas and a disability is not likely, thus no further testing is necessary.*
- *If C-LIM indicates contributory or minimal influence of language/culture, test scores are likely to be valid and the evaluation should continue.*

Step 2. Re-evaluate areas of weakness in native language (L2) to provide additional supporting evidence of validity (cross-linguistic confirmation)

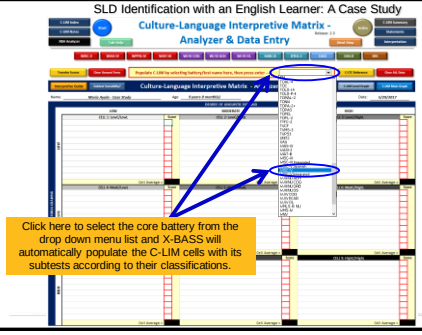
- *If data indicate an area is a strength (i.e., average), then original L2 score is invalid, use the L1 score.*
- *If data indicate an area is still a weakness, then original L2 score is valid, use the L2 score.*

Step 3. Further cross-validate L1 and L2 test scores with contextual factors and pre-referral data and academic concerns (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*

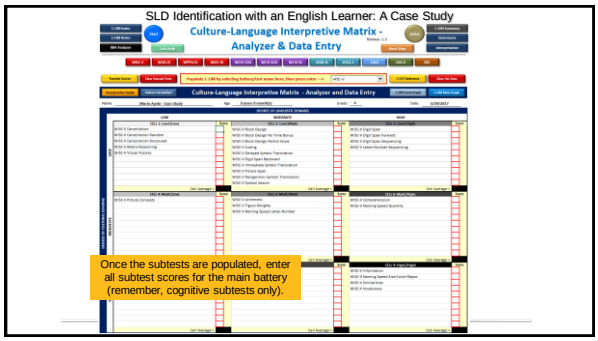
86

SLD Identification with an English Learner: A Case Study

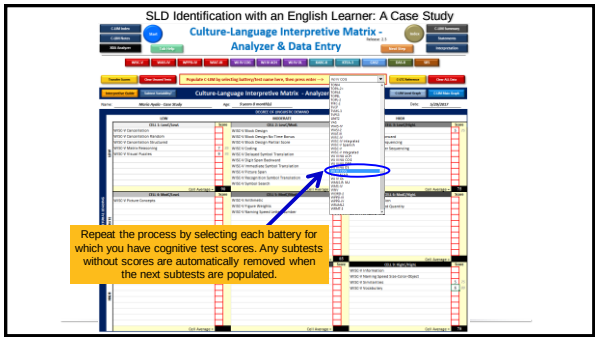


Click here to select the core battery from the drop down menu list and X-BASS will automatically populate the C-LIM cells with its subtests according to their classifications.

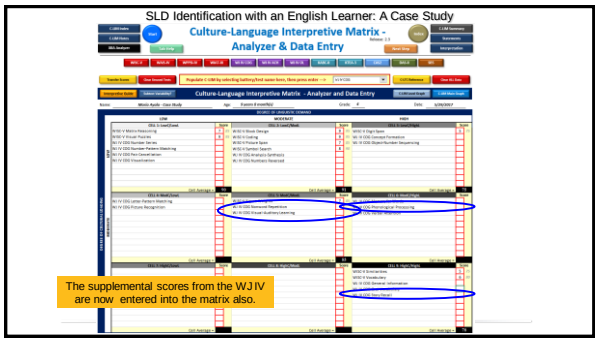
87


SLD Identification with an English Learner: A Case Study
Culture-Language Interpretive Matrix - Analyzer & Data Entry
Once the subtests are populated, enter all subtest scores for the main battery (remember, cognitive subtests only).

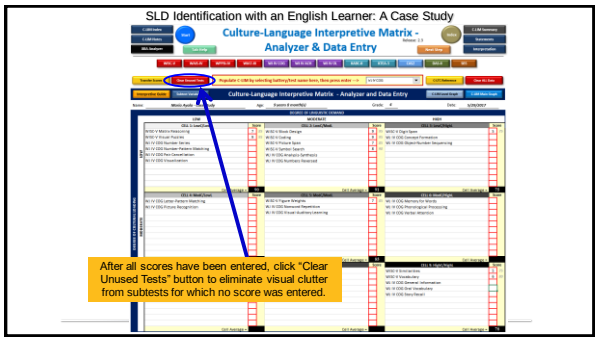
88


SLD Identification with an English Learner: A Case Study
Culture-Language Interpretive Matrix - Analyzer & Data Entry
Repeat the process by selecting each battery for which you have cognitive test scores. Any subtests without scores are automatically removed when the next subtests are populated.

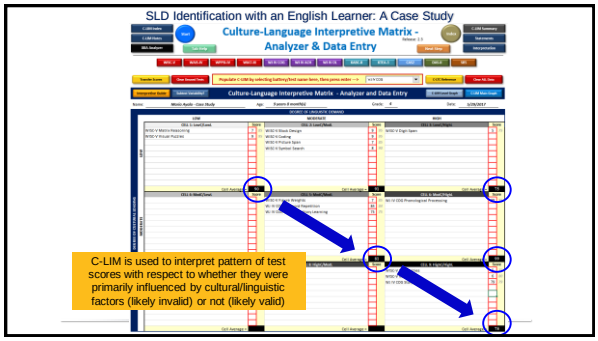
89


SLD Identification with an English Learner: A Case Study
Culture-Language Interpretive Matrix - Analyzer & Data Entry
The supplemental scores from the WJ IV are now entered into the matrix also.

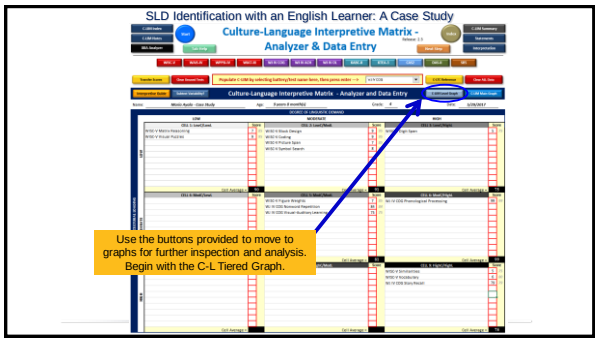
90


SLD Identification with an English Learner: A Case Study
Culture-Language Interpretive Matrix - Analyzer & Data Entry
After all scores have been entered, click "Clear Unused Tests" button to eliminate visual clutter from subtests for which no score was entered.

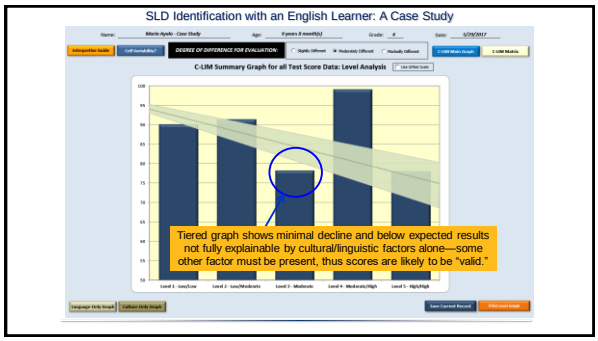
91


SLD Identification with an English Learner: A Case Study
Culture-Language Interpretive Matrix - Analyzer & Data Entry
C-LIM is used to interpret pattern of test scores with respect to whether they were primarily influenced by cultural/linguistic factors (likely invalid) or not (likely valid)

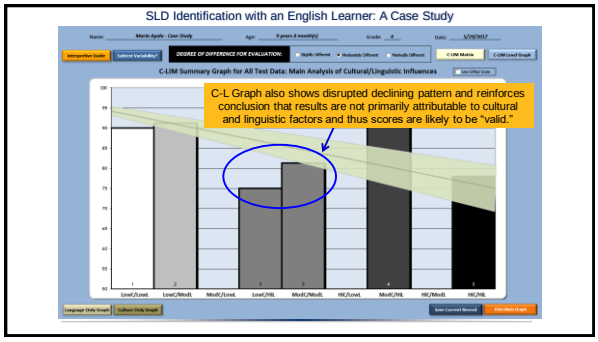
92


SLD Identification with an English Learner: A Case Study
Culture-Language Interpretive Matrix - Analyzer & Data Entry
Use the buttons provided to move to graphs for further inspection and analysis. Begin with the C-L Tiered Graph.

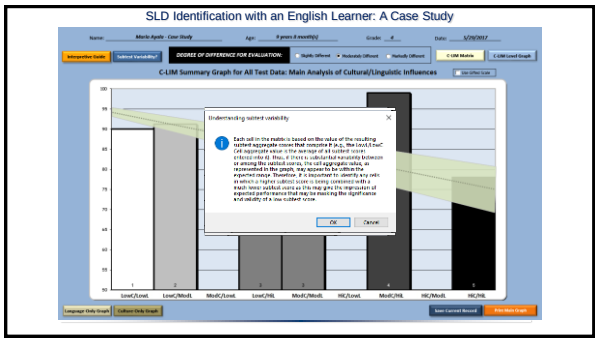
93


SLD Identification with an English Learner: A Case Study
C-LIM Summary Graph for all Test Score Data: Level Analysis
Tiered graph shows minimal decline and below expected results not fully explainable by cultural/linguistic factors alone—some other factor must be present, thus scores are likely to be "valid."

94


SLD Identification with an English Learner: A Case Study
C-LIM Summary Graph for All Test Data: Main Analysis of Cultural/Linguistic Influences
C-L Graph also shows disrupted declining pattern and reinforces conclusion that results are not primarily attributable to cultural and linguistic factors and thus scores are likely to be "valid."

95


SLD Identification with an English Learner: A Case Study
C-LIM Summary Graph for All Test Data: Main Analysis of Cultural/Linguistic Influences

96

Basic Disability Evaluation with an English Learner: A Case Study

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

WECHSLER INTELLIEGENCE 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*
Word Reading	92	Reading Comprehension	76
Pseudoword Decoding	98	Oral Reading Fluency	80

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

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.

97

Basic Disability Evaluation with an English Learner: A Case Study

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

WECHSLER INTELLIEGENCE 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*	
Word Reading	92	Reading Comprehension	
Pseudoword Decoding	98	Oral Reading Fluency	

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

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.

98

Basic Disability Evaluation 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.

99

Basic Disability Evaluation 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:

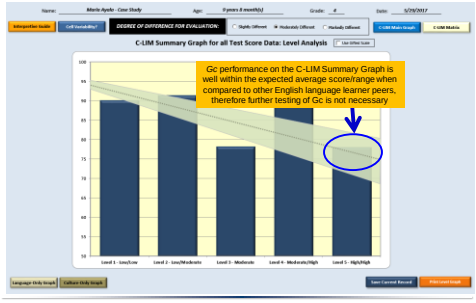
- **Review results from testing in English and identify domains of suspected weakness or difficulty:*
  a. *For Gc only, evaluate weakness according to high/high cell in C-LIM or in context of other data and information*
- **For Gc only:*
  a. *If high/high cell in C-LIM is within/above expected range, 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.*

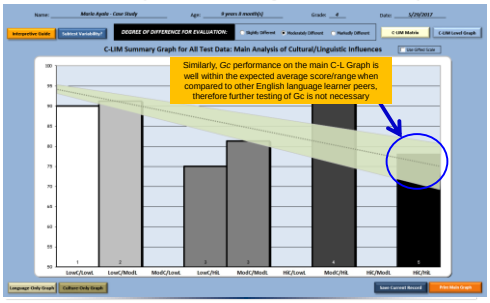
100

Basic Disability Evaluation with an English Learner: A Case Study



101

Basic Disability Evaluation with an English Learner: A Case Study



102

Basic Disability Evaluation 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.

	English	*Native Lang.*	*Valid?*	*Interpretation?*
- Gc	76	-	No	Average
- Gf	82	-	?	?
- Glr	77	-	?	?
- Gsm	78	-	?	?
- Gv	98	-	Yes	Average
- Ga	92	-	Yes	Average
- Gs	94	-	Yes	Average

These are the seven major CHC broad ability or processing areas typically measured for comprehensive evaluation.

Since the aggregate score in the C-LIM for Tier 5 (i.e., the High/High cell where all Gc tests are classified) was within the expected range corresponding to the selected degree of difference deemed most appropriate, it should be considered average despite the fact that the magnitude is only 76 and that it isn't technically a valid measure of intrinsic language-related abilities. This is one reason for the development of the Ortiz PVAT and highlights its utility.

103

Basic Disability Evaluation 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."

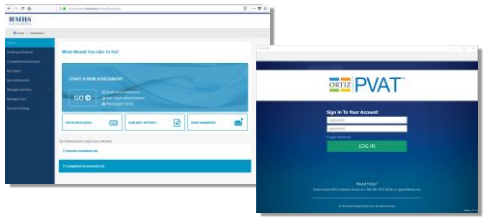
This one issue, more than any other, best highlights the shortcomings of today's tests relative to their failure to provide a true peer comparison group for English learners that would alleviate all of the extra work and potential confusion. There simply is no substitute for being able to make fair and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

104

A Brief Introduction

ORTIZ PVAT™



MHS

105

Fairness and English Learners:
Ensuring True Peer Comparability

Stratification Variables in Dual Standardization Norm Samples of the Ortiz PVAT

English Speakers (*N* = 1,530)

- Ages 2:6 to 22:11
- Gender: equal split
- Stratification:
  - Geographic region
  - Parental education level (PEL)
  - Race/ethnicity

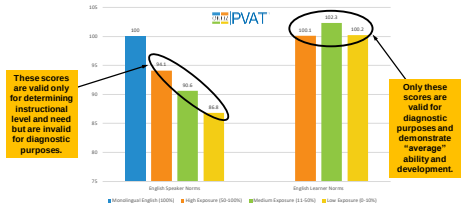
English Learners (*N* = 1,190)

- Ages 2:6 to 22:11
- Gender: equal split
- Stratification:
  - Geographic region
  - Parental education level (PEL)
  - Language spoken at home (53 different languages)
  - Proportion of lifetime exposure to English (i.e., opportunity to learn English):
    - 11 categories for length of exposure to English
    - 0-6 months up to 16+ years

Inclusion of these variables in the stratification of the EL Norm Sample is a completely unique feature of the Ortiz PVAT not found in any other test.

106

The Ortiz PVAT – Advances in fairness and testing
Developmental Language/Exposure-based Comparison Provides Validity and Fairness for ELs



107

The Ortiz PVAT – Fairness for ALL Learners
Removal of all variance due to language results in no influence of race or ethnicity

Norm sample for native English speakers demonstrates negligible effect of race/ethnicity.

Form	Racial/Ethnic Group	*N*	*M*	*SD*	*F (df)*	*p*	Pairwise Comparisons ($p < .01$)	Partial η^2
Form A	Black	280	99.4	15.2	2.60 (3, 1523)	.051	*ns*	.005
	Hispanic	126	99.5	15.4				
	White	1,018	100.5	15.3				
	Other	106	96.3	15.3				
Form B	Black	280	99.6	15.1	2.47 (3, 1523)	.060	*ns*	.005
	Hispanic	126	99.7	15.3				
	White	1,018	100.6	15.2				
	Other	106	96.4	15.2				

108

The Ortiz PVAT – Fairness for ALL English Learners

First language learned (L1) does not alter the sequence of learning English (L2)

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				

109

The Ortiz PVAT – Recommended Applications

Pre-school Screening and Evaluation – dual norms permit evaluation of basic language development (receptive vocabulary) in very young children (minimum age: 2 years, 6 months) in both native English speakers and English learners prior to the beginning of formal instruction.

Progress Monitoring of English Language Proficiency – many tests, for example those used to monitor compliance with Title III ELA requirements are not well designed for that purpose and give misleading results regarding progress and growth and no information relative to the acquisition of BICS vs. CALP.

Determination of Instructional Level – the Assessment Report indicates the linguistically appropriate level of instruction and the degree of intensity required to assist the student in making progress toward grade-level standards and expectations. Specific instructional strategies are also provided.

Progress monitoring of Reading and Writing Vocabulary – the Progress Report provides data for evaluating increases in receptive vocabulary that may reflect relative progress in response to specific interventions that are being employed.

Evaluation of Growth in General Language Ability – unlike tests that do not allow measurement of growth, a specific index documenting actual growth in English vocabulary/language acquisition across short and long intervals is provided.

Development of Intervention/Treatment Strategies – performance is linked directly to specific and customized recommendations for language-based intervention and treatment strategies relative to true peers.

Diagnostic and Disability Evaluation – provides the only norm-referenced "true peer" comparison necessary for evaluating "difference vs. disorder" in general language-related disabilities/disorders related to vocabulary acquisition.

110

Assessment Report from the Ortiz PVAT


ORTIZ PVAT™

Ortiz Picture Vocabulary Acquisition Test™

Samuel O. Ortiz, Ph.D.

Assessment Report

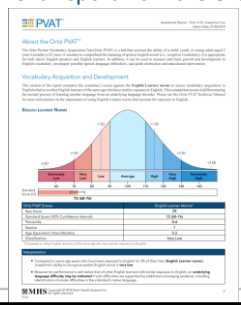
Examinee Information

Assessment Information

MHS

111

Assessment Report from the Ortiz PVAT



112

Performance Across Different Norm Sample Comparisons

How much of a difference does "true language peer" comparison make for diagnostic decisions?

Grade	Age	EL vs. EL: Ortiz PVAT	EL vs. ES: WMLS-III English	EL vs. SS: WMLS-III Spanish
4	9	97	64	40
3	8	87	69	43
4	10	105	63	40
2	7	84	58	42
1	6	98	45	104
5	10	92	42	88
K	5	71	45	40
4	9	97	61	41
4	9	95	55	42
4	9	94	40	61
2	7	92	65	48
1	6	104	68	55
5	9	84	40	73
1	7	89	43	59
	Average=	92	54	56
	Percentile Rank =	30th	0.1st	0.1st
	Potential False Positive Rate =	7-21%	100%	86%

EL = English Learner
ES = English speaker
SS = Spanish speaker

WMLS-III Oral Language
Oral Comprehension
Picture Vocabulary

L1 dominance approach = 12/14 with language impairment

L2 dominance approach = 14/14 with language impairment

True peer comparison = 3/14 with language impairment*

*Of the 3 scores in the true peer comparison, two are very close to being WNL (SEM=2) and may not actually represent a disability.

Without true peer comparison, false positive error rates for misidentification of ELs could be exceptionally high.

Data in this table are provided courtesy of an urban school district and may not be copied or reproduced. Used here with permission of the owner.

113

Basic Disability Evaluation with an English Learner: A Case Study

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

WECHSLER INTELLIGENCE SCALE FOR CHILDREN-V

Verbal Comprehension Index	*76*	*Fluid Reasoning Index*	*82*
Similarities	5	Matrix Reasoning	7
Vocabulary	6	Figure Weights	7
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

Ortiz PVAT (EL Norms) *93*

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.

114

Basic Disability Evaluation 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	Average
- Ga	92	-	Yes	Average
- Gs	94	-	Yes	Average
- Gc (Ortiz PVAT)	93	-	Yes	Average

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.

115

Basic Disability Evaluation with an English Learner: 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 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.*

116

Basic Disability Evaluation 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:

More defensible

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 developmental or dynamic assessment, informal tasks accompanied by careful observation, error analysis, and other probing with the assistance of a translator for communication.*

Less defensible

117

Basic Disability Evaluation with an English Learner: A Case Study

WISC-V/WJ IV/WIAT-III XBA DATA FOR Maria Ayala
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*	*WISC IV Spanish (Gf subtests)*	*91*
Digit Span	5	Coding	9	Matrix Reasoning	8
Picture Span	7	Symbol Search	8	Picture Concepts	9
WISC IV Spanish WMI	*72*				
Digit Span					
Letter-Number Sequencing	4				

WOODCOCK JOHNSON-IV TESTS OF COGNITIVE ABILITY

Auditory Processing	*91*	*LT Storage/Retrieval*	*77*	*Ortiz PVAT*	*93*
Phonological Processing	99	Story Recall	79		
Nonword Repetition	84	Visual-Auditory Learning	75		
		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.

118

Basic Disability Evaluation with an English Learner: A Case Study

WISC-V/WJ IV/WIAT-III XBA DATA FOR Maria Ayala
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*	*WISC IV Spanish (Gf subtests)*	*91*
Digit Span	5	Coding	9	Matrix Reasoning	8
Picture Span	7	Symbol Search	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*	*LT Storage/Retrieval*	*77*	*Ortiz PVAT*	*93*
Phonological Processing	99	Story Recall	79		
Nonword Repetition	84	Visual-Auditory Learning	75		
		Bateria III LT Retrieval	*79*		
		Visual-Auditory Learning	81		
		Retrieval Fluency	78		

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

119

Basic Disability Evaluation with an English Learner: 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.*

120

Basic Disability Evaluation with an English Learner: A Case Study

DETERMINING STRENGTHS AND WEAKNESSES IN MULTILINGUAL EVALUATION

	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*	S	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)	W	S		✓	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)	W	W	✓		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)	S	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.

121

Basic Disability Evaluation 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	Average
- Glr	77	(79)	77 - Yes	Not Average
- Gsm	78	(72)	78 - Yes	Not Average
- Gv	98	-	Yes	Average
- Ga	92	-	Yes	Average
- Gs	94	-	Yes	Average
- Gc (Ortiz PVAT)	93	-	Yes	Average

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.

122

Multilingual Assessment of ELs: Step by Step

Step 1. Test first in English (L2) and evaluate construct validity in all areas in English (exclusion of cultural/linguistic factors)

- *If all scores indicate normative strengths (SS ≈ 90 or higher) when tested in English (L2), scores are valid to the extent that a disability is not likely, thus no further testing is necessary.*
- *If some scores are normative weaknesses (SS < ≈ 90) evaluate test score validity in a research-based manner, e.g., via the C-LIM.*
- *If C-LIM indicates primary influence of language/culture, test scores are likely invalid and indicate average ability in all areas and a disability is not likely, thus no further testing is necessary.*
- *If C-LIM indicates contributory or minimal influence of language/culture, test scores are likely to be valid and the evaluation should continue.*

Step 2. Re-evaluate areas of weakness in native language (L2) to provide additional supporting evidence of validity (cross-linguistic confirmation)

- *If data indicate an area is a strength (i.e., average), then original L2 score is invalid, use the L1 score.*
- *If data indicate an area is still a weakness, then original L2 score is valid, use the L2 score.*

Step 3. Further cross-validate L1 and L2 test scores with contextual factors and pre-referral data and academic concerns (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*

123

Basic Disability Evaluation 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.

124

Basic Disability Evaluation with an English Learner: A Case Study

The Importance of Converging Evidence in Establishing Validity

The student's developmental history relative to culture, language, and education provide the context by which test scores acquire sufficient validity for diagnosing any condition. When test scores are consistent with the referral concerns and the student's experiences, the necessary ecological validity is established for conclusions that suggest the presence of a disability.

	English	*Spanish*	*Valid?*	*Interpretation?*
- Gc	**76**	-	76 - **No**	-
- Gf	(82)	**91**	**91 - Yes**	**Average**
- Glr	**77**	(79)	**77 - Yes**	**Not Average**
- Gsm	**78**	(72)	**78 - Yes**	**Not Average**
- Gv	**98**	-	**Yes**	**Average**
- Ga	**92**	-	**Yes**	**Average**
- Gs	**94**	-	**Yes**	**Average**
- Gc (Ortiz PVAT)	**93**	-	**Yes**	**Average**

To support disability identification on the basis of this pattern of test scores requires consideration of various factors including federal/state regulations and policies, the criteria for whatever approach or model is employed to establish the disability, and especially integration with other data and information that provide a valid, defensible, and consistent picture of the final determination and conclusions. ***Test scores will bolster interpretation only to the extent with which they are consistent with what else is known about the student.***

125

Basic Disability Evaluation with an English Learner: A Case Study

Sample Validity Statement for EL Evaluations

Simplified Validity Statement for LIKELY disability and Determination of VALID Results

Because XXXX is not a native English speaker, it is necessary to establish the validity of test scores to ensure that they are true estimates of their ability and not the result of limited English proficiency.

XXXX's test data were entered into the Culture-Language Interpretive Matrix which permitted evaluation of the extent to which the scores were primarily affected by cultural or linguistic factors. A review of the pattern of test scores indicated that performance *was not consistent* with what would be expected of other individuals with similar cultural and linguistic backgrounds. This means that the scores may be interpreted as fair estimates of XXXX's abilities, with the exception of language which can only be determined to be an area of strength or weakness via comparison to other English learners which was accomplished by further use of the C-LIM.

The statement above is most appropriate for this case where a) the evaluation focused on identification of a suspected cognitive/academic-based disability; and b) where it was determined that the obtained test results were *not* influenced primarily by cultural and linguistic factors, albeit these factors may have remained contributory. Thus, the test results (except for Gc) could be considered valid estimates of the abilities that were measured. Native language testing should also have been conducted to further support cognitive test score validity. This statement has been placed in the public domain and may be freely copied, modified, and distributed for non-profit purposes without the need to secure permission.

126

Basic Disability Evaluation with an English Learner: A Case Study

Sample Validity Statement for EL Evaluations

Simplified Validity Statement for UNLIKELY disability and Determination of INVALID Results

Because XXXX is not a native English speaker, it is necessary to establish the validity of test scores to ensure that they are true estimates of their ability and not the result of limited English proficiency.

XXXX's test data were entered into the Culture-Language Interpretive Matrix which permitted evaluation of the extent to which the scores were primarily affected by cultural or linguistic factors. A review of the pattern of test scores indicated that performance *was consistent* with what would be expected of other individuals with similar cultural and linguistic backgrounds. This means that the scores cannot be interpreted as fair estimates of XXXX's abilities.

However, because the scores were compared to other individuals from research studies who were of average ability and who had not been identified as having a disability, it suggests that XXXX's performance is also average (possibly higher) and that it is not likely that a disability is present in this case. This means that although XXXX is having difficulties in the classroom, the problems are most likely to attributable to, and primarily the result of, the normal process of second language and acculturative knowledge acquisition.

127

Meeting the Standards for Fairness in Evaluation of ELs

Although there are no professional or legal standards that specify actual procedures for evaluation of English learners, there are consensus recommendations that provide some guidance in being able to document and establish that a given evaluation has been conducted in compliance with standards necessary to demonstrate and establish fairness. The following are standards that may be used to assess the extent of fairness and validity of any evaluation.

1. **TOOLS AND PROCEDURES:** The report contains a section detailing the deliberate selection of tools, methods, and procedures with respect to the cultural and linguistic factors in the examinee's background—simply listing tests, even native language ones, is not sufficient. Explanations are provided for any modification or alteration to the administration or scoring of any standardized instrument, including use of a translator or translated test.
2. **DEVELOPMENTAL LANGUAGE HISTORY:** The report report contains a specific and distinct section on language development which contains a detailed history and sufficient information with which to formulate appropriate expectations of current proficiency. Information should include, at a minimum, age of first exposure to all languages, parental/home language use, parental levels of proficiency in all languages, parental education and socio-economic status, individual's experiences with all languages, current proficiency in all languages, amount of formal education in all languages, and type of educational programming.
3. **VALILDITY:** The report contains a section that provides a discussion regarding the validity of the obtained assessment data and test scores including specification regarding how the impact of cultural/linguistic differences were considered and excluded as factors that might have compromised validity of the information—simply stating that scores or data are valid is insufficient.
4. **INTERPRETATION OF RESULTS:** Discussion of results, whether cognitive, linguistic, or academic, are always presented in terms of the extent to which cultural or linguistic factors may have compromised performance and affected interpretive validity and the extent to which they are consistent with or not consistent with what would be reasonably expected of the examinee, given their unique cultural and linguistic background.
5. **DIAGNOSTIC IMPRESSIONS:** The report contains conclusions and interpretations that are supported by integration of data and includes discussion regarding how cultural/linguistic factors are not the primary reasons for any claimed deficits and that such deficits are above and beyond what would be expected given the examinee's unique cultural/linguistic background.

128

Meeting the Standards for Fairness in Evaluation of ELs

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

- generational history
- language proficiency
- socio-economic status
- opportunity to learn
- academic history
- familial history
- developmental data
- work samples
- curriculum based data
- intervention results, etc.

…the framework presented here (along with the C-LIM and Ortiz PVAT) represents an evidence-based method for evaluating English learners and addressing the issue of test score validity. This process can assist **all practitioners** in decreasing the potential for biased and discriminatory interpretation by creating the ability to answer the most basic question in EL assessment:

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

129

Assessment and Related Resources

TESTS:

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

Ortiz PVAT Free 30-Day Trial and 2 Free Uses
http://info.mhs.com/OrtizPVATfreetrial

BOOKS:

Ortiz, S. O., Flanagan, D. P. & Alfonso, V. C. (2015). Cross-Battery Assessment Software System (X-BASS v2.X). New York: Wiley & Sons, Inc.

Ortiz, S. O., Flanagan, D. P. & Alfonso, V. C. (Winter 2019--coming soon). Intervention Library: Finding interventions, resources and supports for students with learning difficulties (IL-FIRST v1.0). New York: Wiley & Sons, Inc.

Flanagan, D. P., Ortiz, S.O. & Alfonso, V.C. (2013). Essentials of Cross-Battery Assessment, Third Edition. 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 XBA Assessment

130