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 assume 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.

"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.

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

Acculturative Language Proficiency – Not Race or Ethnicity

"Most studies compare the performance of students from different ethnic groups... rather than ELI and non-ELI 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 ELI status on observed score differences."

Lohman, Korb & Lakin, 2008

Lohman, Korb & Lakin, 2008
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.

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.

Example of Potential Construct Invalidity:
“Assemble these blocks together in the correct manner so they appear identical to this illustration.”

Example of Potential Interpretive Invalidity:
“After putting a blue block on top of a purple one, put the green block on the blue one.”

Test Score Validity and Defensible Interpretation Requires “True Peer” Comparison

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.

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

Compared to this group, Chaseito’s score is at the 9th percentile rank.

Compared to this group, Panchito’s score is at the 1st percentile rank.

GREEN LINE = Distribution of scores for native Spanish student performance

RED LINE = Distribution of scores for native English student performance

Compared to this group, Chaseito’s score is still likely to be low even if he is receiving L1 instruction.

Compared to this group, Panchito’s score is still likely to be low even if he is receiving L1 instruction.

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.
Compared to this group, Chaseito's score is at the 9th percentile rank. 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.

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

Compared to this group, Panchito's score is at the 1st percentile rank.

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.

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

Compared to this group, Chaseito's score is still likely to be low even if he is receiving L1 instruction.

Compared to this group, Panchito's score is still likely to be low even if he is receiving L1 instruction.

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.
Academic Test Score Validity Requires "True Peer" Comparison

<table>
<thead>
<tr>
<th>Grade Level</th>
<th>Normal Curve Equivalents</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>70(80)*</td>
</tr>
<tr>
<td>5</td>
<td>54(65)*</td>
</tr>
<tr>
<td>4</td>
<td>32(43)*</td>
</tr>
<tr>
<td>3</td>
<td>22(33)*</td>
</tr>
<tr>
<td>2</td>
<td>11(22)*</td>
</tr>
</tbody>
</table>

*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.

The "English only" window

The "achievement gap"

Simultaneous bilingual

Sequential bilingual

Academic Test Score Validity Requires "True Peer" Comparison

Fairness in Evaluation of ELs via RTI/MTSS

The Most Appropriate Standard for Comparison Depends on the Question Being Asked: An RTI example.
Intervention Question: What are Chaseito’s and Panchito’s instructional levels, needs, goals, and how far behind are they academically?

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

Test Score Validity and Defensible Interpretation Requires “True Peer” Comparison

For native English speakers, growth of language-related abilities is 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 classroom 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 comparing frequently used true peers is warranted.” (p. 40)

- Fisher & Frey, 2012
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
- Numerical Methods of Evaluation
  - Avoiding the language by evaluating areas unrelated to language
- Dominant Language Evaluation
  - Choosing a language based simply on relative proficiency

**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 in the testing domain.*
- *any effort to help the examinee perform in the test by circumventing any process involving ‘working around the language’ undermines the test and is not a legitimate method of test administration.*
- *tests are normed to provide valid estimates of examinee abilities using specific test conditions, not ‘working around the language’ strategies.*
- *any of the above actions are violations of standardization and consequently the scores which they provide are not interpretable or defensible.*

*dominant language evaluation,*

- *the violation of the standardized testing process introduces error into the testing process, which cannot be determined to what extent the procedures aided or hindered performance and thus the results cannot be defended.*

**Processes and Procedures for Addressing Test Score Validity**

**ISSUES IN NONVERBAL METHODS OF EVALUATION**

*Language Reduced Assessment: Just avoid the language.*

- *“nonverbal testing” reduces the language barrier, however:*
- *it is impossible to administer a test without some type of communication occurring between examinee and examiner, for example, 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 (Ga and Gc) and mathematics (Gq).*
- *all nonverbal tests are subject to the same problems with norms and cultural content as verbal tests— they do not control for differences in acculturation and language proficiency which may still affect performance.*
ISSUES IN DOMINANT LANGUAGE EVALUATION

1. 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 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.
   - 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 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 speakers.
   - 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 speculative and amount to no more than a guess.

2. Current Approaches Fail to Establish Test Score Validity

<table>
<thead>
<tr>
<th>Evaluation Method</th>
<th>Representative of Bilingual Development</th>
<th>Measures a Wider Range of School-Related Abilities</th>
<th>Does Not Require the Evaluator to Be Bilingual</th>
<th>Adheres to the Test's Standardized Protocol</th>
<th>Substantial Research Base on Bilingual Performance</th>
<th>Sufficient to Identify or Diagnose Disability</th>
<th>Accounts for Variation in Bilingual Development</th>
<th>Likely to Yield Reliable and Valid Data and Information</th>
<th>Provides Extensive Data Regarding Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified or Altered Assessment</td>
<td>✔</td>
<td>✔</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
</tr>
<tr>
<td>Language Reduced Assessment</td>
<td>❌</td>
<td>❌</td>
<td>✔</td>
<td>✔</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
</tr>
<tr>
<td>Dominant Language Assessment in L1: Native Only</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>✔</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
</tr>
<tr>
<td>Dominant Language Assessment in L2: English Only</td>
<td>✓</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✓</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
</tr>
</tbody>
</table>

   All approaches are limited in some manner when addressing test score validity and none are sufficient to diagnose 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.

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

   Whatever method or approach may be employed in evaluation of ELs, 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.
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.


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.

Research Foundations for EL Evaluation

ELs and non-ELs perform differently: Broad ability level

Although it has long been recognized that language likely account for the differences in test performance between English learners 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.
For ELs, tests that require little or no level of age-based acquisition of language and cultural knowledge yield scores at or close to the normative mean.

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

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

Research Foundations for EL Evaluation: EL to ES

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

Fairness in Determining "Average" Performance: ES to ES
Research Foundations for EL Evaluation: EL to ES

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

<table>
<thead>
<tr>
<th>Subtest Name</th>
<th>Scale Score</th>
<th>Scale Score</th>
<th>Scale Score</th>
<th>Scale Score</th>
<th>Grand Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information</td>
<td>7.5</td>
<td>7.8</td>
<td>5.1</td>
<td>7.5</td>
<td>*</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>8.0</td>
<td>8.3</td>
<td>6.1</td>
<td>7.5</td>
<td>87</td>
</tr>
<tr>
<td>Similarities</td>
<td>7.8</td>
<td>8.8</td>
<td>6.6</td>
<td>9.3</td>
<td>89</td>
</tr>
<tr>
<td>Comprehension</td>
<td>7.8</td>
<td>9.0</td>
<td>6.7</td>
<td>8.0</td>
<td>89</td>
</tr>
<tr>
<td>Digit Span</td>
<td>8.3</td>
<td>8.6</td>
<td>7.3</td>
<td>*</td>
<td>89</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>8.7</td>
<td>9.4</td>
<td>7.4</td>
<td>7.8</td>
<td>92</td>
</tr>
<tr>
<td>Picture Arrangement</td>
<td>9.0</td>
<td>10.3</td>
<td>8.0</td>
<td>9.2</td>
<td>96</td>
</tr>
<tr>
<td>Block Design</td>
<td>9.5</td>
<td>10.8</td>
<td>8.0</td>
<td>9.4</td>
<td>97</td>
</tr>
<tr>
<td>Object Assembly</td>
<td>9.6</td>
<td>10.7</td>
<td>8.2</td>
<td>9.3</td>
<td>96</td>
</tr>
<tr>
<td>Picture Completion</td>
<td>9.7</td>
<td>9.9</td>
<td>8.7</td>
<td>9.5</td>
<td>97</td>
</tr>
<tr>
<td>Coding</td>
<td>9.6</td>
<td>10.9</td>
<td>8.9</td>
<td>9.6</td>
<td>97</td>
</tr>
</tbody>
</table>

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

Fairness in Determining “Average” Performance: EL to ES

Typical “average” Range for ELLs

Tests with “low” language demands

Tests with “mod” language demands

Tests with “high” language demands

No decline or below expected range = possible disability

Overall decline and within expected range = no disability
The authors noted that "roughly 97% of (n = 83) of participants were identified as meeting criteria for an educational disability (86% as SLD)" (p. 371). Yet, only 9 ELL cases (10.5%) resulted in invalid scores (no disability). Thus, the C-LIM suggested invalid scores in 9 cases, 3 of which were likely correct (those without disabilities) 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.


**Interpretive Errors in C-LIM Studies: Styck & Watkins**

Overall decline vs. within-expected-range vs. no decline

<table>
<thead>
<tr>
<th>WISC-IV C-LIM Analysis</th>
<th>EL Sample (with disability)</th>
<th>Norm Sample (no disability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invalid Scores (Decline)</td>
<td>N = 9 (N=6, 7.0%/N=3, 3.5%)</td>
<td>N = 100 (4.9%)</td>
</tr>
<tr>
<td>Valid Scores (No decline)</td>
<td>N = 77 (99.5%)</td>
<td>N = 1,933 (95.1%)</td>
</tr>
</tbody>
</table>

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

Research Foundations for EL Evaluation: EL to ES

<table>
<thead>
<tr>
<th>Subtest Name</th>
<th>Mean-EL 1979</th>
<th>Mean-ES 1999</th>
<th>Mean-EL 2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information</td>
<td>S.E.</td>
<td>S.E.</td>
<td>S.E.</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>7.4</td>
<td>7.4</td>
<td>7.2</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>6.9</td>
<td>6.8</td>
<td>6.3</td>
</tr>
<tr>
<td>Reading</td>
<td>7.8</td>
<td>7.6</td>
<td>7.4</td>
</tr>
<tr>
<td>Math</td>
<td>7.0</td>
<td>6.9</td>
<td>6.7</td>
</tr>
<tr>
<td>Science</td>
<td>9.1</td>
<td>8.9</td>
<td>8.6</td>
</tr>
<tr>
<td>Social Studies</td>
<td>8.7</td>
<td>8.6</td>
<td>8.4</td>
</tr>
<tr>
<td>History</td>
<td>9.3</td>
<td>9.1</td>
<td>9.0</td>
</tr>
<tr>
<td>Science</td>
<td>9.5</td>
<td>9.5</td>
<td>9.3</td>
</tr>
<tr>
<td>Social Studies</td>
<td>9.9</td>
<td>9.9</td>
<td>9.7</td>
</tr>
</tbody>
</table>


*Data for this subtest were not reported in the study.*
**Matrix of WISC subtest means arranged by EL vs. ES test performance**

<table>
<thead>
<tr>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Level 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>85</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>Moderate</td>
<td>87</td>
<td>91</td>
<td>97</td>
</tr>
<tr>
<td>High</td>
<td>89</td>
<td>91</td>
<td>97</td>
</tr>
</tbody>
</table>

**Research Foundations for EL Evaluation: EL to EL**

**General ability level performance as compared to other English learners**

Mean WJ III GIA across the four levels of language proficiency on the New York State ESL Achievement Test:

- Proficient: 71.75
- Advanced: 89.55
- Intermediate: 101.0
- Beginner: 70

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

- **Similarities**: 74.76
- **Vocabulary**: 82.29
- **Comprehension**: 69.74
- **Arithmetic**: 73.24
- **Digit Span**: 78.24
- **Block Design**: 75.24
- **Picture Arrangement**: 72.76
- **Picture Completion**: 75.24
- **Coding**: 76.76
- **Object Assembly**: 73.24

**Research Foundations for EL Evaluation: EL to EL**

**Subtest level performance as compared to other English Learners**

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

---

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 age- or grade-expect 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 varied 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 or on comparison to other ELs with the same degree of English exposure and development. 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.

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 it 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 clinicians in making decisions regarding whether test score 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 clinicians in interpretation of test score data in a nondiscriminatory manner.

Free version of C-LIM available at: [http://facpub.stjohns.edu/~ortizs/CLIM/](http://facpub.stjohns.edu/~ortizs/CLIM/)
The Culture-Language Interpretive Matrix – Basic Version 4.0

Degree of Linguistic Demand
Degree of Cultural Loading

Fairness in Determining “Average” Performance: EL to EL
Research-based subtest means regarding expected test performance EL vs. EL
The Culture-Language Interpretive Matrix (C-LIM)

Addressing test score validity for ELLs

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, in 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 (primary influence of cultural and linguistic factors).

4. The principles of ELL test performance as established by research are the foundations upon which the C-LIM is based and serve as a de facto norm sample for the purposes of comparing test results of individual ELLs to the performance of a group of average ELLs with a specific focus on the attenuating influence of cultural and linguistic factors.

The Culture-Language Interpretive Matrix (C-LIM)

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

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

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

When both criteria are observed, it may be concluded that the test scores are likely to have been influenced primarily by the presence of cultural/linguistic variables and therefore are not likely to be valid and should not be interpreted. If either criterion is not met, the results can be assumed to be VALID.

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

A Best Practice Framework for Comprehensive Evaluation of ELs

1. Assess and evaluate factors that affect opportunity to learn and age/grade (foundation level)
   - Identify assessment of first and second language acquisition, type and length of formal schooling.
2. Monitor and evaluate academic skills growth relative to true peers (including native/heritage language or language cautioned evaluation)
3. Assess and evaluate construct validity in all areas in English first (exclusion of cultural/linguistic factors)
4. Testing in both languages (English and native language) is necessary to determine disability
5. All low test scores, whether in English or the native language, must be validated for any conclusions that have been made.
6. Cross-examine the effectiveness of interventions and academic growth. Methods may include:
   - authentic and informal data (e.g., work samples, portfolios, etc.) or more formal data collected within an MTSS/RtI framework (e.g., CBM, progress monitoring charts, standardized test data).

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 performance in the average range is NOT likely a chance finding and strongly suggests average ability (i.e., no deficits in ability)
   - Test performance that is above average (e.g., 90th percentile) is a chance finding because of environmental factors or developmental differences and thus cannot automatically confirm average ability (i.e., possible deficits in ability)
3. Therefore, testing in one language only (English or native language) means that:
   - Is there an indicator of a primary deficit in the other language?
   - Is there any variability in the scores that are suspect?
   - Is there any other evidence that may be masking low performance?
4. Testing in both languages (English and native language) is necessary to determine disability
   - Test performance in both languages is lower than expected compared to true peers, continue evaluation.
   - Test performance in one language is lower than expected compared to true peers, continue evaluation.
5. 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)
6. It CANNOT be determined if the student has a disability (i.e., low scores must be validated as true indicators of deficit ability)
7. 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 without research
Translating Research into Practice

X-BASS v2.3 New Features: PSW Quick Analysis

The PSW Quick Analysis provides a streamlined way to evaluate SLD using only 8 scores (7 cognitive and 1 academic). Although the analysis is nearly 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.

Other cognitive processes may also be entered for 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 neurophys processing domains.

The simplified presentation is easier to comprehend and suitable for printing and inclusion in written reports.

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.
Updated the Test Database with several new tests including: APST, CVLT-II, DTLA-5, EFT (ENG), EFT 3, MBVVT-II, PPVT-R (SP), PPVT-2, RPM-2, RPM-3, RPS-4, TAPS, TAPS-4, TOLD-P-S, TOL 2, TOLD-3:3E, VBAR-5, WISC-V Spanish, and WRAT-5. There are now 148 tests/batteries and 1,175 subtests classified in X-BASS.

X-BASS v2.3 New Features: Test/Battery Updates

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.

X-BASS v2.3 New Features: Two-way PSW Data Transfer

This button will automatically send the selected/listed subtests over to their respective domains in the OBA-Analyzer according to their primary CHC broad ability classifications.
X-BASS v2.3 New Features: Selectable/Other Ability Domain

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.

X-BASS v2.3 New Features: Separation of Glr into Gl and Gr

Gl (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.

X-BASS v2.3 New Features: Exclusionary Factors Form Tab

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.
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.

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.

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.
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.

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.

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.
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.

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 weaknesses without the influence of unrelated abilities.

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.
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.

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

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.

Additional guidance is available to assist in interpreting C-LIM results within the matrix.
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.

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.

Assistance is available for evaluating score variability that may mask true weaknesses between tiers in the C-L Level Graph.
Additional guidance is available to assist in interpreting C-LIM results in the C-L Main Graph.

Assistance is available for evaluating subtest variability within cells that may mask true weaknesses in the C-L Level Graph.

New, simplified validity statements for use with the C-LIM are provided alongside the previous detailed statements. These may be more helpful in explaining procedures, results, and interpretation within written reports in comparison to the more detailed and technical versions.
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

DDE: 6/22/2016
DOB: 10/4/2006
Grade: 4

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) when tested in English (L2), scores are valid to the extent that a disability is not likely. Thus, no further testing is necessary.
• If scores are normative weaknesses (SS < 90), evaluate test score validity in a research-based manner, e.g., via the C-LLIM.
• If C-LLIM indicates a primary influence of language/culture, test scores are likely invalid and indicate average ability in all areas and a disability is not likely. Thus, further testing is unnecessary.
• If C-LLIM indicates a 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, ensures consistent with current standards, and permits AH.flex testing to occur. And in some cases, completes the testing without having bilingual or requiring outside assistance. However, C-LLIM may require knowledge of research to determine whether test scores are valid, and if any group comparison is appropriate.
Basic Disability Evaluation with an English Learner: A Case Study

WECHSLER INTELLIGENCE SCALE FOR CHILDREN - V

<table>
<thead>
<tr>
<th>Index</th>
<th>Verbal Comprehension Index</th>
<th>Fluid Reasoning Index</th>
<th>Visual-Spatial Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>WISC-IV</td>
<td>76</td>
<td>82</td>
<td>95</td>
</tr>
</tbody>
</table>

Similarities: 5
Matrix Reasoning: 7
Block Design: 9
Vocabulary: 6
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

<table>
<thead>
<tr>
<th>Achievement Area</th>
<th>Basic Reading</th>
<th>Reading Comprehension</th>
<th>Written Expression</th>
<th>Word Reading</th>
<th>Reading Comprehension</th>
<th>Spelling</th>
<th>Pseudoword Decoding</th>
<th>Oral Reading Fluency</th>
<th>Sentence Composition</th>
<th>Essay Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>94</td>
<td>76</td>
<td>92</td>
<td>92</td>
<td>92</td>
<td>100</td>
<td>98</td>
<td>86</td>
<td>93</td>
<td>93</td>
</tr>
</tbody>
</table>

WOODCOCK JOHNSON

<table>
<thead>
<tr>
<th>Test</th>
<th>Auditory Processing</th>
<th>LT Storage/Retrieval</th>
<th>Phonological Processing</th>
<th>Story Recall</th>
<th>Nonword Repetition</th>
<th>Visual-Auditory Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>91</td>
<td>77</td>
<td>99</td>
<td>79</td>
<td>85</td>
<td>75</td>
</tr>
</tbody>
</table>

WISC - V/WJ IV/WIAT

XBA DATA FOR Maria Ayala


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 = 89.0 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 < 89.0) 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.

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.
Once the subtests are populated, enter all subtest scores for the main battery (remember, cognitive subtests only).

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.

The supplemental scores from the WJ IV are now entered into the matrix also.
After all scores have been entered, click "Clear Unused Tests" button to eliminate visual clutter from subtests for which no score was entered.

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).

Use the buttons provided to move to graphs for further inspection and analysis. Begin with the C-L-Tiered Graph.
Trend 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."

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."

E-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."
Basic Disability Evaluation with an English Learner: A Case Study

There are four possible causes of a genuine weakness that may suggest itself as an area in the required academic difficulties as well as three areas of strength. However, these procedures are not designed for linguistically confirm that they are 'deficient' as three areas of strength. Therefore, practitioners must find and rely on a "true peer" comparison group such as that comprised primarily of native English speakers. Moreover, testing in the native language cannot be compared fairly to native English speakers (it doesn't).

In addition, because Gc itself is "language," it is deemed "valid" using the C-M at this stage, additional procedures are necessary to determine whether it is a strength or weakness even when scores are deemed "valid" using the C-M. Thus, in the case, additional procedures are necessary to determine whether it is a true weakness or not and whether it has or does not require remediation.

Because Gc, by definition, comprised of cultural knowledge and language development, the influence of these factors cannot be separated from tasks designed to measure them. Thus, when comparing to English or a controlled variable of a non-English 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-M 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 ELLs as being the same (they aren't), as being behind in English (a temporary, not final), and 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 of the C-M to help ensure that ELLs are not unfairly regarded as having either deficient Gc ability or significantly lower overall cognitive ability that ELLs are not unfairly regarded as having either deficient Gc ability or significantly lower overall cognitive ability.

There are four possible causes of a genuine weakness that may suggest itself as an area in the required academic difficulties as well as three areas of strength. However, these procedures are not designed for linguistically confirm that they are 'deficient' as three areas of strength. Therefore, practitioners must find and rely on a "true peer" comparison group such as that comprised primarily of native English speakers (it doesn't).

In addition, because Gc itself is "language," it is deemed "valid" using the C-M at this stage, additional procedures are necessary to determine whether it is a strength or weakness even when scores are deemed "valid" using the C-M. Thus, in the case, additional procedures are necessary to determine whether it is a true weakness or not and whether it has or does not require remediation.

Because Gc, by definition, comprised of cultural knowledge and language development, the influence of these factors cannot be separated from tasks designed to measure them. Thus, when comparing to English or a controlled variable of a non-English 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-M 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 ELLs as being the same (they aren't), as being behind in English (a temporary, not final), and 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 of the C-M to help ensure that ELLs are not unfairly regarded as having either deficient Gc ability or significantly lower overall cognitive ability—that may simultaneously decrease identification of SLD and increase suspicion of ID and speech impairment.
Re-evaluation of suspected areas of weakness is necessary to provide cross-linguistic confirmation of evidence of disability. Cross-linguistic assessment is important as it helps determine if the observed difficulties are specific to one language or if they occur across languages. 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.
- For Gc only, evaluate weaknesses according to high/high cell in C-LIM or in context of other data and information.
- 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).
- 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.

Basic Disability Evaluation with an English Learner: A Case Study

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

<table>
<thead>
<tr>
<th>English</th>
<th>Native Long</th>
<th>Valid?</th>
<th>Interpretation?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gc</td>
<td>76</td>
<td>No</td>
<td>Average</td>
</tr>
<tr>
<td>Gf</td>
<td>82</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Gl</td>
<td>77</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Gsm</td>
<td>77</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Gv</td>
<td>98</td>
<td>Yes</td>
<td>Average</td>
</tr>
<tr>
<td>Ga</td>
<td>92</td>
<td>Yes</td>
<td>Average</td>
</tr>
<tr>
<td>Gs</td>
<td>94</td>
<td>Yes</td>
<td>Average</td>
</tr>
</tbody>
</table>

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

Interpretive Problems with Gc Scores with English Learners

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.

Basic Disability Evaluation with an English Learner: A Case Study

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 allow for a more comprehensive understanding of their potential and provide a subscore 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 provides valid subscores for English learners of any language background and level of English exposure—and that test is the Ortiz PVAT.
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.

Stratification Variables in Dual Standardization Norm Samples of the Ortiz PVAT

Fairness and English Learners: Ensuring True Peer Comparability

This graph is reproduced from the Technical Manual of the Ortiz PVAT and is Copyright © 2017 Multi-Health Systems Inc. All rights reserved.

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

These scores are valid only for determining instructional level and need but are invalid for diagnostic purposes.

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

These scores are valid only for diagnostic purposes and demonstrate “average” ability and development.

The Ortiz PVAT – Fairness for ALL Learners

Removal of all variance due to language results in no influence of race or ethnicity

This table is reproduced from the Technical Manual of the Ortiz PVAT and is Copyright © 2017 Multi-Health Systems Inc. All rights reserved.
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**

<table>
<thead>
<tr>
<th>Form</th>
<th>Language Spoken</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>F (df)</th>
<th>P</th>
<th>Partial Comparisons (q-r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form A</td>
<td>Spanish &amp; Spanish Creole</td>
<td>67</td>
<td>71.4</td>
<td>8.6</td>
<td>1.66</td>
<td>.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intra-European Languages</td>
<td>151</td>
<td>78.2</td>
<td>6.5</td>
<td>4.36</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asian &amp; Pacific Islander Languages</td>
<td>120</td>
<td>79.3</td>
<td>8.4</td>
<td>0.12</td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>All Other Languages</td>
<td>25</td>
<td>79.7</td>
<td>9.4</td>
<td>2.93</td>
<td>.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form B</td>
<td>Spanish &amp; Spanish Creole</td>
<td>87</td>
<td>71.4</td>
<td>8.6</td>
<td>1.66</td>
<td>.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intra-European Languages</td>
<td>134</td>
<td>78.2</td>
<td>6.5</td>
<td>4.36</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asian &amp; Pacific Islander Languages</td>
<td>120</td>
<td>79.3</td>
<td>8.4</td>
<td>0.12</td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>All Other Languages</td>
<td>25</td>
<td>79.7</td>
<td>9.4</td>
<td>2.93</td>
<td>.09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This table is reproduced from the Technical Manual of the Ortiz PVAT and is Copyright © 2017 Multi-Health Systems Inc. All rights reserved.

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**

The Ortiz PVAT – Recommended Applications

**Pre-school Language and Evaluation— dual norm parallel evaluation of basic language development (receptive vocabulary) in very young children (minimum age: 2 years, 6 months) in both native English speakers and young 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 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**—only 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.
Performance Across Different Norm Sample Comparisons
How much of a difference does "true language peer" comparison make for diagnostic decisions?

<table>
<thead>
<tr>
<th>Grade</th>
<th>Age</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
<th>9th</th>
<th>10th</th>
<th>11th</th>
<th>12th</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**True Peer Comparison**
- 12/14 with language impairment
- 14/14 with language impairment
- 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.

Potential False Positive Rate =
- 7 - 21% 100% 86%

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

WECHSLER INTELLIGENCE SCALE FOR CHILDREN
- 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

WECHSLER INDIVIDUAL ACHIEVEMENT TEST
- Basic Reading 94
- Reading Comprehension 76
- Written Expression 92
- Word Reading 92
- Reading Comprehension 76
- Spelling 100
- Pseudoword Decoding 98

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

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 since we don't have any individual to generate or to identify language impairment in the case of an EL. We also note that, in this case, an easily generated summary report would provide the information in an interpretive summary report.

Basic Disability Evaluation with an English Learner: A Case Study
Basic Disability Evaluation with an English Learner: A Case Study

Avoiding Interpretive Problems by Use of the Ortiz PVAT

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 on the C-LIM and scores below the expected range of difference in performance may be discounted in the light of factors other than cultural knowledge or language ability. However, using the English Learner norms in cross-cultural/linguistic situations may be defensible in cases where the range of performance may well be deemed to be the result of factors other than cultural knowledge or language ability. However, to establish validity for a low score obtained from testing in English with an ELL, other than Gc, appropriate for determining areas of suspected weakness using tests administered in English for abilities other than Gc.

As mentioned previously, 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 defensible to use the Ortiz PVAT score because it was derived precisely on “true peers” and therefore inherently valid in terms of discerning “learning ability” from “learning disability.”

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 on the C-LIM and scores below the expected range of difference in performance may be discounted in the light of factors other than cultural knowledge or language ability. However, using the English Learner norms in cross-cultural/linguistic situations may be defensible in cases where the range of performance may well be deemed to be the result of factors other than cultural knowledge or language ability. However, to establish validity for a low score obtained from testing in English with an ELL, other than Gc, appropriate for determining areas of suspected weakness using tests administered in English for abilities other than Gc.

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 on the C-LIM and scores below the expected range of difference in performance may be discounted in the light of factors other than cultural knowledge or language ability. However, using the English Learner norms in cross-cultural/linguistic situations may be defensible in cases where the range of performance may well be deemed to be the result of factors other than cultural knowledge or language ability. However, to establish validity for a low score obtained from testing in English with an ELL, other than Gc, appropriate for determining areas of suspected weakness using tests administered in English for abilities other than Gc.

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
3. Use of English language tests translated directly by a bilingual evaluator
4. Use of English language tests translated directly by a bilingual evaluator
5. Use of English language tests translated directly by a bilingual evaluator
6. Use of English language tests translated directly by a bilingual evaluator

These are the guidelines from the best practice recommendations apply to all abilities, including Gc, when it has been determined to be a weakness because it falls below the expected range of differences in the C-LIM.

Avoiding Interpretive Problems by Use of the Ortiz PVAT

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 on the C-LIM and scores below the expected range of difference in performance may be discounted in the light of factors other than cultural knowledge or language ability. However, using the English Learner norms in cross-cultural/linguistic situations may be defensible in cases where the range of performance may well be deemed to be the result of factors other than cultural knowledge or language ability. However, to establish validity for a low score obtained from testing in English with an ELL, other than Gc, appropriate for determining areas of suspected weakness using tests administered in English for abilities other than Gc.
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 in an alternate language testing is based on a score in the alternate language, it serves to validate/confirm the corresponding weakness score obtained in English or lower scores in the alternate language to provide additional confirmation that they are true weaknesses. The scores in some tests can be used and scores may be generated for the main purpose of the alternate performance as part of the report to provide cross-language information of acceptance of validity.

- 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 serve to bolster the validity of the original score obtained in English.

Based on these guidelines, 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 PsBA 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 validate/confirm the corresponding weakness score obtained in English or lower scores in the native language to provide additional confirmation that they are true weaknesses.

- 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 invalidate/disconfirm the corresponding weakness score obtained in English.

- For all domains, including Gc, if a score obtained in the native language also suggests weakness in the same domain (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 serve to bolster the validity of the original score obtained in English. This approach is recommended in cases of low family SES, or other lack of opportunity to learn the native language, in cases of native language difficulties where a native language score is needed for a behavioral, educational, or medical evaluation, or when using alternate language testing for DL, dual, and SL.

- For all domains, including Gc, if a score obtained in the native language also suggests weakness in Gc (SS < 90), it may serve to bolster the validity of the original score obtained in English.

- For all domains, including Gc, if a score obtained in the native language also suggests weakness in the same domain (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 serve to bolster the validity of the original score obtained in English. This approach is recommended in cases of low family SES, or other lack of opportunity to learn the native language, in cases of native language difficulties where a native language score is needed for a behavioral, educational, or medical evaluation, or when using alternate language testing for DL, dual, and SL.

- For all domains, including Gc, if a score obtained in the native language also suggests weakness in the same domain (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 serve to bolster the validity of the original score obtained in English. This approach is recommended in cases of low family SES, or other lack of opportunity to learn the native language, in cases of native language difficulties where a native language score is needed for a behavioral, educational, or medical evaluation, or when using alternate language testing for DL, dual, and SL.
### Basic Disability Evaluation with an English Learner: A Case Study

**Determining Strengths and Weaknesses in Multilingual Evaluation**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Native Language</th>
<th>English</th>
<th>Low?</th>
<th>Not Average</th>
<th>Average</th>
<th>High?</th>
<th>Interpreted?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gc</td>
<td>76</td>
<td>75</td>
<td>No</td>
<td>Yes</td>
<td>Average</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gf</td>
<td>90</td>
<td>92</td>
<td>Yes</td>
<td>No</td>
<td>Average</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gv</td>
<td>86</td>
<td>82</td>
<td>No</td>
<td>Yes</td>
<td>Average</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glr</td>
<td>92</td>
<td>90</td>
<td>Yes</td>
<td>No</td>
<td>Average</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gsm</td>
<td>84</td>
<td>80</td>
<td>Yes</td>
<td>No</td>
<td>Average</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Although the 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 the expected range in C. For ALL domains, LIM indicates contributory or minimal influence of language/culture. Test scores are likely invalid and indicate low scores in both native and non-native languages. 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.*

### 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 CLILM.
- If there is a difference 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 CLILM 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. Follow cross-validated 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 conclusion.
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 basic forms of evidence that bolster the validity of obtained test scores by any expectation of test performance that may have to be learned or taught. The evaluation of test scores may also include information from other forms of evidence (e.g., 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.

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

1. Test results from evaluation in English that have been subjected to systematic analysis of the influence of cultural and linguistic factors and against which performance has been compared to other normed or criterion-referenced standards.
2. Test results on other tests requiring production of skills in a native language that either further support or contradict evidence of performance on the test or the specific characteristics of the test due to evidence of cultural or language dominance.
3. Ecological and contextual evidence regarding consistency of the test scores with ecological data and information on development and influence (e.g., L1 and L2 exposure, language of presentation, socio-economic status, parental education level, etc.) and convergence of patterns of performance with other case data (e.g., progress monitoring data, parent reports, teacher/licensed reports, observation, school records, 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.

Basic Disability Evaluation with an English Learner: A Case Study

The Importance of Converging Evidence in Establishing Validity

The evidence considered in this section includes the student’s developmental history relative to culture, language, and education; parental education level; socio-economic status; economic status; parental education level; socio-economic status; economic status; and evidence of disability and determination of disability. The Importance of Converging Evidence in Establishing Validity

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. The 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.

The statement above is most appropriate for this case where a) the evaluation focused on identification of a suspected cognitive/academic disability; and b) where it was determined that the obtained test scores from evaluation in English were likely to be influenced by cultural and linguistic factors.

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. The 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.

The statement above is most appropriate for this case where a) the evaluation focused on identification of a suspected cognitive/academic disability; and b) where it was determined that the obtained test scores from evaluation in English were likely to be influenced by cultural and linguistic factors.
Results with what would be TOOLS AND PROCEDURES: DEVELOPMENTAL LANGUAGE HISTORY: DIAGNOSTIC IMPRESSIONS: language proficiency INTERPRETATION OF RESULTS: generational history developmental data VALIDITY: academic history opportunity to learn socio intervention results, etc. familial history curriculum based data work samples acculturative knowledge acquisition.

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 attributable to, and primarily the result of, the normal process of second language and acculturative knowledge acquisition.
Assessment and Related Resources

BOOKS:

ONLINE:
- CHC Cross-Battery Online
- Free C-LIM Resources
- Ortiz Picture Vocabulary Acquisition Test (Ortiz PVAT) http://info.mhs.com/OrtizPVATfreetrial
- Ortiz PVAT Free 30-Day Trial and 2 Free Uses http://info.mhs.com/OrtizPVATfreetrial

TESTS:
- Ortiz Picture Vocabulary Acquisition Test (Ortiz PVAT) http://info.mhs.com/OrtizPVATfreetrial
- Ortiz/PVAT Free 30-Day Trial and 2 Free Uses

Competency-based XBA Certification Program
- CHC Cross-Battery Online
- Ortiz Picture Vocabulary Acquisition Test (Ortiz PVAT) http://info.mhs.com/OrtizPVATfreetrial
- Ortiz PVAT Free 30-Day Trial and 2 Free Uses

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

Assessment and Related Resources