

Chapter 13: Procedures for Conducting Intellectual Assessment of English Learners

Developed in collaboration with Dr. Samuel O. Ortiz, St. John’s University, New York

Contents

Introduction	196
Goals of Nondiscriminatory Assessment	197
Comprehensive Assessment	199
Processes and Procedures for Addressing Test Score Validity	200
Philosophy of Nondiscriminatory Intellectual Assessment	203
Framework for Nondiscriminatory Assessment of Cognitive Abilities	204
Instructions for Culture-Language Interpretive Matrix (C-LIM) Use, Analysis, and Interpretation	218
Additional Considerations Regarding the C-LTC and C-LIM	231
Frequently Asked Questions Regarding Intellectual Assessment and the C-LIM	232
Resources	235
Tools	238

Note: This manual is meant to be advisory only and does not constitute legal advice or represent an official legal position of the Department of Education. School Districts and individuals are responsible for compliance with state and federal law. Any contrary statements or incorrect information in agency manuals do not negate the provisions of law.

Introduction

One of the most complex tasks in the evaluation spectrum involves nondiscriminatory assessment of the cognitive, linguistic, and academic abilities of individuals who are culturally and linguistically diverse. The failure to assess the abilities of diverse individuals in an equitable manner has been identified as one of the main reasons for the disproportionate representation of various ethnic and linguistic minority groups in special education. Because of the centrality of assessment to the identification of disabilities, such over-representation indicates that at least some, perhaps many, individuals have been mistakenly identified as having a disability when in fact they do not. Of course, avoiding equitable assessment, either because of the complexities or lack of knowledge, may well lead to under-representation of individuals from diverse backgrounds who do in fact have a disability.

For those who work with school-aged children, federal law, as contained in the Individuals with Disabilities Education Act (IDEA 2004), provides some of the most important regulatory requirements and procedural specifications related to nondiscriminatory assessment and governs the evaluation of students who are suspected of having a disability. IDEA 2004 continues to require the longstanding mandate that “assessments and other evaluation materials used to assess a child under this part- Are selected and administered so as not to be discriminatory on a racial or cultural basis” (§300.304(c)(1)(i)). While the intent here is noble, the reality is that such a requirement is primarily aspirational because absolutely strict enforcement would eliminate the use of nearly every tool or procedure currently employed in such evaluation. Compared to the wide range of assessments for students who are native English speakers and raised in U.S. mainstream culture, there is a conspicuous absence of both tools and methods that are not discriminatory in some manner. For whatever reasons, research and development regarding nondiscriminatory evaluation practices and instruments has remained rather stagnant and has not moved much beyond simplistic notions regarding translation and adaptation of existing tests into other languages. The greatest obstacle to nondiscriminatory assessment, however, lies more in the extent to which the evaluator possess the requisite knowledge and competency to engage in equitable practices despite such limitations. It is the evaluator who actually represents the most potentially biased aspect of any evaluation. This remains true even in cases where the evaluator is culturally and linguistically diverse. As stated by Flanagan, McGrew and Ortiz (2000), “mere possession of the capacity to communicate in an individual’s native language does not ensure appropriate, nondiscriminatory assessment of that individual. Traditional assessment practices and their inherent biases can be easily replicated in any number of languages” (p. 291). In addition, given the increasing diversity in the U.S., nondiscriminatory assessment has already become an unavoidable challenge for every evaluator in every corner of the U.S. which means that it must, therefore, embody a set of practices and procedures accessible to everyone, regardless of training, experience, or background.

Goals of Nondiscriminatory Assessment

In the broadest sense, nondiscriminatory assessment is a process designed to ensure fairness in evaluation procedures designed to promote unbiased decisions that result in equitable outcomes regardless of the individual’s ethnicity, language, background and experiences. It is important to understand that the focus of nondiscriminatory assessment rests on the issue of fairness and equity and should not be seen solely as reliance on particular methods intended to promote more racial balance in special education. In this sense, true nondiscriminatory assessment is really something that can be applied with all children not just those who are culturally and linguistically diverse. Practitioners are advised to engage in these practices because they result in better evaluations and consequently better decisions about educational programming, not because they meet legal requirements or change the ethnic composition of children in special education. With respect to individuals who are culturally and linguistically diverse, the major hurdle in the process centers on addressing and answering what seems to be

a simple question: **To what extent do the student’s observed educational difficulties reflect a cultural or linguistic difference vs. a disorder?**

Note that when considering the use of tests and other measures of ability, this relatively straightforward question is in fact tied to the concept of construct validity. If it can be established that the test data obtained in an evaluation of an individual are valid, that is, the scores accurately estimate the individual’s true level of the actual abilities that were measured, irrespective of cultural/linguistic factors, then such scores may be readily interpreted in the same manner as is usually accomplished. Conversely, if it is established that the test data obtained in an evaluation of an individual are not valid (or validity cannot be determined), it means the scores cannot be viewed as either reliable or valid measures of the intended abilities (due to the confounding influence of cultural/linguistic factors) and therefore cannot be interpreted. Only when it is determined that test scores or any collected data are valid can a diagnosis of disability be rendered.

It is possible that the concept of validity has long been misunderstood and this is why the question of “difference vs. disorder” persists without any easy answers. It is equally plausible, however, that assessment of diverse individuals has remained hampered by a range of factors including the lack of specific professional standards for evaluation which permits chaotic and unsystematic practice, inertia regarding unproven clinical recommendations and procedures, and the illusion of the availability of appropriate and valid tools and tests. It is not uncommon for assessment professionals faced with conducting evaluations of diverse individuals to view the process merely as a search for the “right” tool or the “best” test. When combined with inadequate or a lack of training in pre-service programs, it is easy to understand why this occurs. Moreover, given the obvious problems that arise as a function of language differences that affect and inhibit effective communication (as opposed to cultural differences that are not nearly as well understood), most of the focus by publishers in creating evaluation instruments has centered almost exclusively on overcoming the perceived language barrier—notably by creating tests in languages other than English. Some of this impetus no doubt emanates from two sources. One involves the mistaken notion that one can generate valid results simply by testing in an individual’s dominant language. Such a view ignores the fact that an individual may well be dominant, that is, more proficient, in one language than another but that does not guarantee that the individual possesses truly age-expected developmental proficiency in that language. For example, an English learner may both prefer to use English and may test as being English dominant after only 3-4 years of formal education beginning in kindergarten. It would be incorrect to assume that such “dominance” translates to valid test scores when evaluated in English. Likewise, such assumptions ignore the fact that in the U.S., an English learner is, and always will be, bilingual, not monolingual, and that being evaluated as if they speak only one language is an egregious error. This point is given even greater significance when it is considered that the regulations in IDEA permit evaluation of an English learner who has been re-designated from limited English proficient (LEP) to fluent English speaker, as if they were a monolingual native English speaker with no consideration of their permanent bilingual status.

The second source stems directly from the regulatory requirements in IDEA that govern evaluation procedures. IDEA 2004 requires that assessments “Are provided and administered in the child’s native language or other mode of communication and in the form most likely to yield accurate information on what the child knows and can do academically, developmentally, and functionally, unless it is clearly not feasible to so provide or administer.” (§300.304(c)(1)(ii) emphasis added). This mandate actually expands the old provision and even places emphasis on the collection of data that are most likely to yield valid data which recognizes that native language evaluation may not be sufficient to accomplish this goal. Nonetheless, the direct wording regarding the use of the “native language” often misdirects evaluation efforts toward native language assessment as the primary, and often only, strategy for conducting a fair evaluation. Language is only one part of the problem in evaluation, and while it has a central place of importance, there are other aspects regarding fairness of evaluation that must be paid at least equal attention. These aspects include the notion that native language evaluation is not always the most accurate or even feasible approach.

Comprehensive Assessment

The process of conducting fair and equitable assessments is without question multi-faceted and there continues to be a dearth of structural guidelines for practice that offer a comprehensive framework for nondiscriminatory evaluation. One example of such a framework was published initially by Ortiz (2002), and has been updated since that time. The most current revision was published in 2014. This approach consists of 10 basic steps as outlined in Table 10.1.

The framework for Ortiz (2002) makes it clear that nondiscriminatory assessment is more than selecting the “right” test or providing native language evaluation. In addition, the emphasis is placed on working in a systematic manner because bias reduction is accomplished only when actions are taken in an appropriate way and in an appropriate sequence. When attempts to reduce the discriminatory aspects of evaluation are marred by random and haphazard modifications or changes in the normal evaluative process, the results cannot be readily evaluated and quickly lose their meaning and significance. Although the focus of this chapter is on cognitive and intellectual assessment, particularly the use of standardized tests in the course of such evaluations, practitioners are advised to remember that testing forms only one small part of the overall framework for conducting nondiscriminatory assessment.

Table 10.1. A Best Practice Framework for Nondiscriminatory Assessment

1.	Assess for the purpose of intervention.
2.	Assess initially with authentic and alternative procedures.
3.	Assess and evaluate the learning ecology.
4.	Assess and evaluate language proficiency and development.
5.	Assess and evaluate opportunity for learning.
6.	Assess and evaluate relevant cultural and linguistic factors.
7.	Evaluate, revise, and re-test hypotheses.
8.	Determine the need for and language(s) of formal assessment.
9.	Reduce potential bias in traditional assessment practices.
10.	Support conclusions via data convergence and multiple indicators.

Table adapted from: Ortiz, S.O. (2014). Best Practices in Nondiscriminatory Assessment. In P. Harrison & A. Thomas (Eds.) Best Practices in School Psychology VI: Foundations (pp. 61-74), Bethesda, MD: National Association of School Psychologists.

Processes and Procedures for Addressing Test Score Validity

The focus of this chapter is on cognitive and intellectual assessment, particularly as accomplished via the use of standardized tests. Thus, it is Step 9, “Reduce potential bias in traditional assessment practices” listed in the framework for best practice outlined in Table 10.1 that is of central concern to the present discussion. Historically, there have been three general approaches to address validity which require examination regarding the degree to which there is any research to support the validity of their use. That is, do any of the traditional methods employed in evaluation of English learners actually produce results that are valid and permit interpretation? As will be discussed, there are advantages and disadvantages to all three methods but ultimately, none of them are fully satisfactory in producing valid test score data.

1. Modification and Alteration in Assessment: Perhaps the most common modification to testing is the use of a translator/interpreter for administration. The idea is that using one helps to overcome the language barrier and thus might produce valid results. Unfortunately, unless a test has been standardized via use of a translator/interpreter (none have been thus far), using one violates the standardization protocol of the test and effectively undermines score validity, even when the interpreter is highly trained and experienced. Another example of modified and altered assessment involves efforts to help the examinee perform to the best of his/her ability. This process is sometimes referred to as “testing the limits” and may include such procedures as alteration or elimination of specific 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. However, as with the use of a translator/interpreter, employing any of these methods violates the standardization protocol even when “permitted” by the test publisher, unless separate norms for such altered administration are provided (none are). Thus, 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. This is not to say that such alterations shouldn’t be done, but more to note that their use is perhaps most helpful in deriving qualitative information—observing behavior, evaluating learning propensity, evaluating developmental capabilities, analyzing errors, etc. As such, a recommended procedure would be to administer tests in a standardized manner first, which would then potentially allow for later interpretation (e.g., if they are validated by use of the C-LIM), and then consider any modifications or alterations that will further inform the referral questions. Overall, 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.

2. Language Reduced Assessment: Generally known as “nonverbal testing,” the use of language-reduced or “nonverbal” tests are, on the surface, quite helpful in overcoming the language obstacle. However, in point of fact, it is actually impossible to administer a test without some type of communication occurring between examinee and examiner. Although some such tests employ pantomime for administration, the teaching of the meaning and the use of such nonverbal gestures represents the creation of a common language, albeit a non-oral language, that is used for the purposes of communication to complete the testing. Thus, the idea that nonverbal testing is language free is quite misleading as communication, in whatever form, remains an absolute necessity for the purposes of testing. Beyond this, a review of language-reduced tests indicates a heavy and not surprising reliance on visual stimuli, some of it real objects. Thus, some language-reduced tests remain very culturally embedded—they do not become culture-free simply because language is not required for responding. A more important problem for language-reduced tests is related to construct underrepresentation. Because language-reduced tests deliberately exclude language-related abilities, there is a de facto narrowing of the abilities that are measureable in this manner. Most such tests provide measures of fluid reasoning (Gf), visual processing (Gv), and visual short-term memory (Gsm). This can be a major issue when an evaluation is centered on language-related abilities, such as might occur in evaluation of specific learning disability in the areas of reading and writing. Within the context of CHC theory, the two most important cognitive abilities related to a potential learning disability in reading or writing are crystallized knowledge (Gc) and auditory processing (Ga). Neither of these abilities are measured by language-reduced tests. Moreover, all language-reduced tests are subject to the same problems with norms and cultural content as verbal tests—that is, they do not control for differences in acculturative knowledge acquisition and developmental language proficiency which may still affect performance, albeit less than with verbal tests. In sum, language-reduced tests are helpful in the evaluation of diverse individuals and may provide better estimates of true functioning in certain areas that have little

or no relationship to language. 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.

3. Native Language (L1) Assessment: Often referred to as “bilingual assessment,” this procedure actually refers to a native language assessment of an English learner by a bilingual professional who has determined that the examinee is more proficient (i.e., “dominant”) in their native language than in English. The origin of this idea may have come from an old edition of the Standards for Educational and Psychological Testing (AERA, APA & NCME, 1999). One specification contained in the volume suggested that individuals should be tested in their dominant language. While this prescription recognizes that performance in the dominant language is likely to be better than performance in the non-dominant language, it does not in any way ensure that the obtained results are valid. Nevertheless, the practice persists and there has been little awareness that being “dominant” in the native language does not imply age-appropriate development in that language or that formal instruction, and that it is unlikely that all formal education has been only in the native language, or that both the development of and formal instruction in that language have remained uninterrupted by any other. The moment a student enters the U.S. educational system, they must learn English and they automatically become circumstantial bilinguals, and therefore not comparable to monolinguals of either language. Another potential problem with this approach is that although the bilingual psychologist is able to conduct assessment activities in the native language, this option is not directly available to the monolingual psychologist. Given the relative lack of appropriately trained and highly qualified bilingual psychologists, it is impossible to promote an approach that can only be accomplished by a tiny fraction of such uniquely competent professionals. Even were there more bilingual professionals, the very concept of native language assessment is a relatively new idea and an unexplored research area. There is very little empirical support to guide appropriate assessment activities or upon which to base standards of practice or evaluate test performance fairly. And finally, whether a test permits evaluation only in the native language or in some combination of the native language and English (i.e., presumably “bilingual”), the norm samples they contain continue to provide inadequate representation due to the lack of attention to developmental language proficiency and variable opportunity for acculturative knowledge acquisition. In short, there is certainly value to be had by testing an individual in their native language. However, bilinguals in the U.S. are not the same as monolinguals elsewhere and without a research base on which to compare performance of bilinguals, there is simply no way to evaluate the validity of the obtained test results using this approach only.

It is clear that all three typical approaches in use today for evaluating English learners fail in the same regard with respect to producing valid test scores. Each method has at its core, an attempt to reduce what are perceived as threats to test score validity particular as related to language. And each method can provide important information but the value of it will be more along qualitative lines than quantitative ones. True nondiscriminatory assessment will likely rely on

some aspects of these approaches but there is a significant need for steps that fully and directly address the question of validity. Without being able to establish test score validity, any subsequent interpretation or attempt to assign meaning becomes little more than a simple guess. Speculation of this kind is completely indefensible and does not in any way embrace the concept of evidence-based practice and should be avoided at all cost. Rather, it will be the ironic fact that English learners have been tested in English for over a century that provides a viable route for nondiscriminatory assessment despite being the very reason why these other approaches were developed in the first place.

Philosophy of Nondiscriminatory Intellectual Assessment

The process of assessment conducted on culturally and linguistically diverse students should be based on a philosophy that respects and responds to the various idiosyncratic cultural and linguistic factors involved in each case. Such a philosophy might include the principles described in Table 10.2.

Table 10.2. Principles of Best Practices in Conducting Psychoeducational Assessment

1. **Focus on the fundamental question.** All efforts and activities conducted in regard to the process of assessment, including pre-referral activities, should seek to answer the fundamental question which is “why is the student unable to learn normally within the context of the regular classroom?”
2. **Use an hypothesis driven process.** It is important to begin the referral and evaluation process by exploring the hypothesis that the causes of the individual’s learning difficulties are due to external factors. That is, assessment is conducted with the notion that there is nothing wrong with the individual and that systemic, ecological, or environmental factors are the primary reason for the observed learning problems. This hypothesis is maintained until the collected data and information suggest otherwise and when all plausible external factors are ruled out.
3. **Conduct focused assessment.** Assessments should be focused and targeted toward the gathering of data to answer specific questions and hypotheses. Assessment should not be exploratory in nature and needlessly broad or vague. Assessment should not be conducted in a manner that seeks to uncover whatever dysfunction might arise by chance. Assessment is like a hunting trip that targets specific game, not a fishing trip that casts a wide net to see what might be pulled in spontaneously. In addition, procedures such as observation and interview should also be focused and targeted at the gathering of information that is relevant to answering the specific questions and referral concerns.
4. **No “standard battery.”** Assessments should be designed to meet the particular circumstances and idiosyncrasies of the referral concerns in each individual case. As such, the use of a “standard battery” in assessment is an unacceptable method of practice and violates the legal mandates outlined in IDEA 2004 which specify that assessments be appropriate and individualized. The student’s cultural background and linguistic history must be used to form the context within which assessment efforts are selected, conducted and ultimately against which data are interpreted.

5. **No routine testing.** Testing may or may not be a part of any assessment. Particularly, the use of standardized tests is not always a necessary component of every assessment and should not be used on a routine basis. When standardized tests are deemed necessary, only those tests necessary to answer specific questions and hypotheses should be given. The administration of unnecessary or superfluous tests should be avoided.
6. **Conduct systematic assessment.** Assessment should be systematic, logical, and guided by an established framework that is specifically designed to evaluate the areas of functioning that are relevant to the referral questions. Where necessary or relevant, assessment activities should be conducted in a manner that reduces any potential bias or discrimination to the maximum extent possible.
7. **Consider all data as important.** Data gathered from activities other than testing such as from observations, interviews, record reviews, authentic assessment, work samples, and so forth, are to be given equal weight in the determination of the causes of an individual's presumed learning difficulties. Test data are neither more "objective" nor more important than other types of data.
8. **Use multiple, corroborating data sources.** Decisions regarding the primary cause of the individual's problems cannot be based on only one procedure, method, source, or type of data. There must always be corroborating evidence among multiple sources and types of data in order to support any conclusions that are proffered.
9. **Link assessment to intervention.** It is the intent of any assessment to provide information that can resolve the learning issues. Thus, the end result is not to diagnose so much as it is to intervene. Data are therefore gathered in order to identify interventions that will be most helpful to the student. Data should be used to make modifications to a student's educational program and promote more success in the classroom. Likewise, reports that document the assessment should describe the process outlined above and answer the fundamental question regarding the cause of the individual's learning difficulties. The report reflects the professional opinion of the assessor and includes at least a clear description and summary of the findings, a diagnosis (if applicable), and specific recommendations for intervention.
10. **Recognize your limits of competency.** Practitioners should be aware of their own limits of competency related to assessment of diverse individuals and should not engage in any activities for which they do not have sufficient training or expertise. Others with the requisite expertise should be consulted as may be necessary in the course of any evaluation.

Framework for Nondiscriminatory Assessment of Cognitive Abilities

As noted previously, any framework for nondiscriminatory assessment must be accessible to all practitioners, not merely a subset of them with special skills or competency. This point was reinforced by the publication of the first position statement issued by the National Association of School Psychologists that references bilingualism. The paper, "The Provision of School Psychology Services to Bilingual Students" (NASP, 2015) provides the very first attempt by NASP to provide a policy to guide evaluation efforts with English learners. It is important to note that the term "bilingual" in the title of the position paper is used to modify the word "students" and not "school psychology." The reason is to reinforce the idea that such policy binds all school psychologists and that the focus of the guidelines is on what should be done when working with students from diverse linguistic backgrounds so as to improve and increase the fairness of such

evaluations. Such a position highlights the presence of certain variables that must necessarily affect and guide recommendations for practice. Indeed, when considering all such factors, there is a natural logic and flow to the creation of a framework that can be employed by all practitioners rather than just a few.

Consider for example that within the educational venue, the most common purpose for testing is to identify a particular disability that might qualify and entitle a student to special education services. And by definition, the identification of a disability is predicated on the presence of low, not average or higher, test scores. Scores must also be valid before they can be interpreted but because the focus of disability evaluations is predicated upon low scores, validity is less of a concern for ability scores that are average or higher. This is because it is extremely unlikely that an individual would be able to “guess” their way to an average or higher score on the tasks that are typically used in psychoeducational evaluations. Thus, whereas it may be true that an average or higher score may not be valid, it is only invalid in the sense that it could be underestimating the individual’s true ability. But because it is known that the individual’s ability cannot be any less than average, this automatically rules out support for the presence of a disability. The same cannot be said of low scores, however, because they may well be invalid due to numerous other confounding variables (e.g., inattention, lack of motivation, incorrect scoring or administration, etc.). In this sense, the low score doesn’t represent lack of actual ability as much as it reflects the attenuating effect of a confounding variable. What this means is that testing in a single language (either English or the native language) can produce results that could rule out a disability as would be the case if all scores were average or higher. Because this is a rare occurrence in testing that follows a suitable pre-referral process, the reality is that there may well be at least one low score suggestive of a deficit in any given evaluation. At this point, nondiscriminatory evaluation requires that the low scores be validated as true estimates of the individual’s ability (or lack thereof, in this case). It is here that the process of evaluation hits a crossroad. If the evaluation was initially conducted via the native language, there is a sudden absence of procedures or methods by which to assess the validity of the obtained low scores. Given the problems inherent in native language tools, particularly around the issue of the lack of “true peer” norm samples, there is simply no way to determine whether such scores are valid estimates of poor ability. Conversely, if the evaluation was conducted initially in English, there is a vast body of literature upon which to assess the degree to which cultural and linguistic factors may have affected test performance. Indeed, this is the sole purpose of the Culture-Language Interpretive Matrix. Because the only way to determine whether obtained test scores are valid and permit interpretation is via use of test administered in English, it makes sense that any best practice recommendation embrace a process which begins in this manner. In addition, doing so means that all practitioners, regardless of the student’s linguistic background, can initiate any evaluation without assistance and remain engaged in a process that would have otherwise excluded them completely had it started with native language evaluation instead.

A final consideration in the development of a nondiscriminatory framework involves acknowledgement that a disability cannot be identified solely from evaluation in one language as a true disorder must exist in both languages. Thus, even evaluations that initially begin in English must therefore incorporate some type of native language evaluation to provide cross-linguistic and confirmatory evidence to support the presence of a disability that was first identified via testing in English. Given these practical, logical, and psychometric issues, the following framework outlines a general process for planning and carrying out fair and equitable evaluation of cognitive abilities that can be performed, at least to some extent, by any practitioner and which attends to the central question of validity. It should be pointed out that there may be some instances where this framework may not actually be the most ideal or appropriate. However, for the vast majority of psychoeducational evaluations and given the skills and limits of competency of the general practitioner, this framework serves as an appropriate guide under most circumstances. The steps are as follows:

1. Review existing information on the student's language background, language proficiency, culture, and educational history to provide the proper context for test score interpretation. Collect additional information if needed using the tools and questions found in Chapters 6 and 7.
2. Develop an appropriate battery which best addresses the referral concerns and which responds to the requirements necessary for identifying any facet of the suspected disability.
3. Test in English first and use the C-LIM to evaluate test score validity. If the scores are deemed to be invalid, or if all test scores indicate strengths (average or higher) a disability is not likely and thus no further testing is necessary.
4. If the scores are deemed to be valid via use of the C-LIM, and if some scores from testing in English indicate weaknesses, re-evaluate those areas in the native language to support them as areas of true weakness.
5. Use multiple indicators and converging evidence to support the ecological validity of all decisions and conclusions.

1. Review and collect background information

Formal assessment of cognitive abilities is not the first step in the evaluation process. As described in Chapter 6, the team needs to engage in a series of data gathering efforts before using standardized tests. The information to be sought prior to the evaluation of cognitive abilities is crucial in setting the context for interpreting results fairly. Of the various types of information to be collected, perhaps the most important are those which relate to the student's opportunity for the acquisition of cultural knowledge and the amount of exposure to English (not just conversational, but also in terms of advanced language capabilities as compared to native speakers). The purpose of gathering such data is to determine how "different" the individual is from the mainstream along these two dimensions and to create the proper context by which to formulate appropriate and reasonable expectations of test performance.

In general, the curriculum, the teacher's training, the administration, the classroom environment, the expectations, the methods for monitoring progress, and everything else related to the school as a system is designed to allow learning to take place in children who are

from the “mainstream” and otherwise normal. Those kids who seem not to benefit from instruction are thus “different” than those kids who do and merit special programming and educational assistance. But this works only when all students are comparable and have the same level of experience with schools, the same language, and so forth. Thus, children who are culturally and linguistically diverse may not demonstrate expected levels of learning in this system, not because they are not capable, but because they are “different.” Therefore, the extent to which it can be stated validly that poor performance in school learning is due to some intrinsic factor of the student has to do with the degree to which all other sources of the problem have been eliminated or controlled.

Therefore, in the preliminary stages of the referral and assessment process, the focus rests on understanding how “different” the student is compared to the average, mainstream, monolingual English speaking student for whom all these processes and procedures and instruction and intervention have been designed. Naturally, the more “different” the student is, the more it would be expected that poor performance is a function of this difference and not some internal problem. Conversely, the more similar a student is to the mainstream, the more likely that repeated failure to respond to appropriate instruction is due to some internal dysfunction. Knowledge of the degree of the student’s differences on the dimensions of English proficiency and acculturative knowledge acquisition not only assists in understanding the student’s response to instruction, but also sets the level of expectation for performance on any task that may be given, including standardized tests should the matter go that far.

Determining a student’s level of language proficiency is relatively straightforward. In Minnesota, students identified as English learners (ELs) are given English proficiency screening tests when they enter an EL program, as well as annual statewide ELP assessments. There are also many standardized tests that can be used to gauge language development, such as the Woodcock-Munoz Language Survey, the Language Assessment Scale (LAS), and the newly developed Ortiz Picture Vocabulary Acquisition Test (Ortiz PVAT). The key here is not to overestimate development, particularly by paying attention to surface aspects of speech such as pronunciation or the presence of an accent. Accent is not an indicator of language proficiency so much as it is an indication regarding when an individual first began to learn the language. Any individual under the age of about 9 or 10 will likely be able to learn how to pronounce English in a year or two so that they might be mistaken for having the same level of proficiency as their native-English speaking peers. Table 10.3 provides a summary of myths related to language acquisition that can assist practitioners in avoiding assumptions about proficiency and development that may not be true or representative of the individuals they may be assessing.

Table 10.3. Language Acquisition Myths

Myth	Reality
Accent is an indicator of proficiency.	No, it is a marker regarding when an individual first began to hear/learn the language.
Children learn languages faster and better than adults do.	No, they only seem to because they have better pronunciation.
Language development can be accelerated.	No, but having developed one language to a high degree (CALP) does help in learning a second language more easily.
Learning two languages leads to a kind of linguistic confusion.	No, there is no evidence that learning two or more languages simultaneously produces any interference.
Learning two languages leads to poor academic performance.	No, on the contrary, students who learn two languages very well (CALP in both) tend to outperform their monolingual peers in school.
Code-switching is an example of a language disorder and poor grammatical ability.	No, it is only an example of how bilinguals use whatever words may be necessary to communicate their thoughts as precisely as possible, irrespective of the language.

As has been discussed, the two major factors that affect test performance include the student’s current level of developmental proficiency (particularly in English) and the degree to which they have been given an opportunity to acquire the cultural knowledge and content that finds its way onto cognitive ability tests. There is, of course, a relationship between acculturation, language proficiency, and a family’s immigration history. Dr. Catherine Collier describes just such a relationship as part of the research and development behind the “Acculturation Quick Screen” (AQS). The AQS asks eight questions about the length of time in the U.S., the length of time in the district, first and second language proficiency, and characteristics of the current school. Based upon the answers to these questions, students are classified as follows:

- Significantly less acculturated: at the beginning stages of adapting to the current school environment.
- Less acculturated: the student is in the process of adapting but may be experiencing stress and anxiety as a result.
- In transition: the student is in the midst of the acculturation process and still experiencing some culture-shock.
- More acculturated: the student still needs some support, but is generally able to understand and function in the new environment.
- Highly acculturated: the student is able to understand and function in the school environment without support; the student may need encouragement to maintain ties to their traditional cultural community.

Dr. Collier recommends cultural interventions related to the stage of acculturation, and also recommends that this information be used when planning special education assessment strategies.

Of particular note is the distinction between “acculturation” and “acculturative knowledge acquisition.” The former term generally refers to the degree to which individuals identify with a particular culture. The latter only represents the amount of time that an individual has been exposed to a given culture. The AQS provides a measure of this latter dimension and it should be understood that even children who are or describe themselves as being high acculturated still cannot be compared fairly to monolingual, mainstream English speakers who have been in the process of acculturation themselves for their entire lives. As such, English learners are never completely acculturated because they never have the same amount of opportunity to develop the acculturative knowledge that accompanies those who have been acquiring it their entire lives. In many ways, this process mirrors that of language development in that when the point comes where an English learner is both dominant in English and fully proficient (at least by school and district standards) they are still not comparable to monolingual English speakers who have been learning English only for their entire lives. Again, the main purpose of activities within this step are to provide information that can be used to determine how “different” the student is from the mainstream. It is this difference that sets up appropriate expectations for performance on tests and which constitutes the true peer reference standard against whom the child’s performance is to be compared. For the purposes of evaluations within this framework, the degree of difference can be gauged simply as “slightly different,” “different,” or “markedly different.” Caution should be used not to **overestimate** an individual’s English language proficiency or the opportunity they may have had for the acquisition of cultural knowledge. A simple maxim that may assist in heeding this caution states, “once a bilingual, always a bilingual.”

2. Develop an Appropriate Battery that Responds to the Referral Concerns and Evaluates All Facets of the Suspected Disability

It is important that psychologists understand that any evaluation in which there is a suspected disability must be designed in a manner that will provide the greatest chance at identifying said disability. The need to properly evaluate any given disability necessitates that the decisions regarding the most appropriate test for a particular age, which ability areas are and are not measured by various batteries, and which are the most responsive to the nature of the referral concerns and the suspected disability cannot be summarily over-ridden by linguistic or cultural considerations. The factors that may confound test performance relative to language or acculturative knowledge acquisition can be evaluated and ruled out. Thus, if the initial battery compiled for a given evaluation was modified or constrained by cultural and linguistic considerations, the resulting data may be insufficient to address the very nature of the referral concerns or serve the purpose of disability identification. Consequently, there should be somewhat of a balance in making such decisions, but ultimately, preference should be given to

ensuring that should the results be deemed valid, that they are sufficient for the purposes of the evaluation otherwise additional and follow up testing may be required.

As an example, it was pointed out earlier that language-reduced tests tend to include measures of abilities that can be easily evaluated via language-reduced tasks. Thus, it is not uncommon for such tests to measure aspects of functioning and ability that are very amenable to abstract stimuli and novel tasks that can avoid language use or relevance as much as possible. This means that abilities such as visual processing (Gv), fluid reasoning (Gf), and visual short-term memory (Gsm) are commonly measured by the tasks found on language-reduced tests. Such abilities, however, may be insufficient for the purposes of some evaluations where the nature of the disability may be rooted in difficulties that are manifested in reading and writing skills. Therefore, if an evaluation for specific learning disability is planned for a student who is displaying reading and writing difficulties, and only language-reduced tests are used, it would greatly reduce the likelihood that the evaluation uncovers the processing ability that explains the observed academic skills deficits as they would most likely stem from problems in crystallized intelligence (Gc) or auditory processing (Ga). Thus, the decision to only use a language-reduced test for evaluation of specific learning disability in an English learner results in a narrowing of abilities measured such that it would make it exceptionally difficult to uncover and pinpoint the true nature of the child's learning difficulties. Similar problems may occur in cases where insufficient attention is paid to the theoretical organization of a battery and the breadth of abilities measured. For these reasons, it is recommended that psychologists consider adopting methods that can ensure psychometric and theoretical validity even before the issue of construct validity related to linguistic/cultural factors can be considered as is embodied, for example, in the Cross-Battery Assessment (XBA) approach (Flanagan, Ortiz & Alfonso, 2013).

3. Test in English first and use the C-LIM to evaluate test score validity.

The recommendation here is predicated upon use of the C-LIM. To use the C-LIM, any tests that were selected as being appropriate to the referral, must have been administered strictly according to the standardization protocol and in English only. The C-LIM cannot be used in cases where any alteration or modification of tests has been utilized as such procedures introduce error in ways that cannot be identified or determined and thus prohibit evaluation by the C-LIM which is based on research conducted on English learners of average ability and who are not disabled.

The sole purpose of the C-LIM is to evaluate the degree to which cultural and linguistic influences may have affected test scores, to the point that they may not be valid representations of the individual's ability, and therefore cannot be interpreted. More importantly, the C-LIM is not a test or scale of any kind and should not be thought of as such. While it may be correct to call it a tool or an instrument, the C-LIM does not in any way "measure" something. Rather, the C-LIM simply takes research, calculates the expected performance of English learners from the mean values drawn from research, and then organizes it in a simple way for practitioners to be able to compare the performance of their examinee to

the known and expected average performance of ELs drawn from research studies. Note that this is only necessary because there are no tests (with the exception of the Ortiz PVAT) that has actual norm samples that constitute "true peers" and which permit fair comparisons. Thus, the only way in which a practitioner can comply with the legal, procedural, and psychometric requirements regarding the validity of obtained test scores (as a function of cultural and linguistic factors) is to create as close to a true peer comparison group as possible and research studies are the only way to do that.

In this sense the C-LIM is in fact merely a guide to practice, not a tool or an instrument that measures any ability. That is, the C-LIM is designed to assist practitioners in using research to inform their decisions (i.e., whether their results are valid or not) so that they are engaging in practice that can be called evidence-based. Consider that without using the C-LIM, a practitioner would find it difficult, if not impossible, to claim that their test scores are valid. How a practitioner could establish that their results are valid and on what basis such a claim could be made without use of the C-LIM remains a mystery. Proper use of the C-LIM is based on organizing existing research findings in a way that facilitates determination regarding whether cultural and linguistic influences affected the obtained test results in a clear, overall, and systematic manner. If the pattern of results demonstrates a distinct pattern of decline across the entirety of the scores, and when the magnitude of the scores is within the expected range associated with the degree of difference possessed by the individual, then the results can be determined to be likely invalid and thus cannot be used to represent the actual level of ability for the individual.

If testing has been conducted in English, and if the C-LIM results indicate that the results are likely to be invalid (because there is an overall pattern of decline and the bars in the graphs are within the expected/shaded range in the C-LIM), then further testing becomes unnecessary. This is because sufficient data have been obtained to permit the conclusion that the examinee does not have a disability because the pattern of scores is consistent with and in the same range as that of other English learners with similar cultural/linguistic backgrounds who are not disabled and are of average ability or higher. This determination is precisely the same as what is typically done with native English speakers wherein a graph with shading is used to depict the average range as determined by the norm sample. The graph represents various levels of functioning and when the performance of the individual is mapped onto it, it provides a norm referenced standard by which comparative performance can be classified accordingly. Again, this means then that you can be sure the examinee does not have a cognitive-based disability, albeit there is uncertainty regarding what the actual level of their abilities might be. Nevertheless, the data provide conclusive evidence that the individual has no cognitive deficiencies. As noted previously, this information is sufficient for the purposes of a disability evaluation because in the absence of any scores that are lower than the "average" or expected range of performance, it cannot be reasoned that there is any impairment or disorder. Moreover, if the examinee doesn't display any deficits at all in English, then by definition they should not have any deficits in their native language either as it's not possible to be disabled in one but not the other. And because

low scores (deficits) are more of a concern when it comes to validity, the presence of nothing but average scores means there are likely no deficits because an average score on an ability test cannot generally be ascribed to luck or chance. A low score may well be due to a factor other than lack of ability but an average or higher score is most certainly an indication of at least average ability. Thus, in a disability evaluation determining that an examinee does NOT have any deficits serves the same purpose as saying that the examinee doesn't have a disability because deficits are required to identify a disability.

Conversely, in cases where the pattern of test scores as viewed within the C-LIM do not reveal an overall pattern of decline or where there is at least one cell in the matrix that is lower than what would be expected given the individual's degree of difference, it may be reasonably concluded that the results are likely to be valid—that is, they represent valid estimates of the individual's true ability (or lack thereof). At this point, the identification of possible deficits in one or more areas of cognitive functioning must be further validated as it is not sufficient or appropriate to base a disability solely on test results obtained in English. There must be evidence of cross-linguistic confirmation and validity and this process is outlined in the next step.

4. Re-evaluate Areas of Weakness via Native Language

Whenever analysis via the C-LIM indicates that the results are likely to be valid (due to lack of an overall pattern of decline or because the scores within the cells were lower than the expected range based on degree of difference), the next step involves cross-linguistic validation and confirmation of the validity or existence of any areas that have been identified as weaknesses via testing in English. Because true deficits cannot exist in only one language, it is imperative that any presumed deficits identified from initial testing in English be re-evaluated in the native language as well.

Thus, while it can be determined that an individual does **not** have a disability by testing in English first (which would occur only if the examinee showed no deficits when tested in English—an admittedly rare occurrence), it **cannot** be determined that an examinee has a disability unless evaluation is completed in both languages. In this step, however, it is equally important to recognize that there is no need to re-evaluate areas in which the individual displayed average or higher functioning when tested initially in English. Doing so would be a serious waste of resources and time and would do nothing to inform the already known fact that the individual does not have a deficit in that area. Within this framework, only areas that are indicative of deficits should be re-evaluated but there are differences between how possible deficits in Gc should be handled as opposed to deficits in all other areas.

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.

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 weakness according to high/high cell in C-LIM or in context of other data and information.
- For Gc only:
 - 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 if and when required, albeit with appropriate nondiscriminatory assignment of meaning, and that it be used for the purposes of instructional planning and educational intervention. Doing so, however, may 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 that is considered average and a processing “strength” (i.e., it facilitates learning). This one issue with Gc, 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.

In response to the many difficulties posed by these issues, a new test has been developed with dual-norm samples, including one specifically for English learners that yields valid Gc scores for

English learners of any language background and level of English exposure—and that test is the Ortiz PVAT. The most unique feature of the Ortiz PVAT is the incorporation of dual norm samples (one specifically for native English speakers and one specifically for English learners) as well as the inclusion of an additional and critical stratification variable in the EL norm sample to control for differential amounts of English language exposure. In this way, an English learner can be compared to other English learners of the same age and with the same amount of exposure and opportunity for learning English, regardless of their native language. As such, the Ortiz PVAT is capable of providing standard scores for diagnostic purposes for both native English speakers and English learners that do not require examination of validity within the C-LIM and scores that may be interpreted in the usual manner without fear of attenuation from linguistic/cultural influences. And it does so in the area which is most problematic in evaluation (Gc) and appears to be a viable alternative for measuring this ability as compared to existing measures (e.g., WISC-V VCI). The test represents perhaps a new and pioneering direction in test development and is an example of the way in which standardized tests can be improved in a manner that helps to simplify evaluation by addressing the concept of validity as related to differential experiences in learning English and acculturative knowledge acquisition.

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 EL 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:
 - For all abilities, except Gc, evaluate weakness using standard classifications (e.g., SS < 90).
- Re-test all domains of suspected weakness, including Gc when it is not within the expected range of difference in the C-LIM* using native language tests.
- Administer native language tests or conduct re-testing using one of the following methods:
 - Native language test administered in the native language (e.g., WJ III/Bateria III or WISC-IV/WISC-IV Spanish).
 - Native language test administered via assistance of a trained interpreter.
 - English language test translated and administered via assistance of a trained interpreter.

- Administer tests in manner necessary to ensure full comprehension including use of any modifications and alterations necessary to reduce barriers to performance, while documenting approach to tasks, errors in responding, and behavior during testing, and analyze scores both quantitatively and qualitatively to confirm and validate areas as true weaknesses.

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.” Thus, there should be no concern about incidental exposure to tests as this would be insufficient to increase performance much or at all in cases where a true disability is present.

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

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

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

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

Because average or higher scores in testing are unlikely to be due to chance, when a score obtained from native language testing in this step 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

5. Use Multiple Indicators and Converging Evidence to Support the Ecological Validity of All Decisions and Conclusions.

Apart from test score data generated and evaluated for validity in the preceding steps, it is necessary to obtain additional, converging evidence to further substantiate and provide ecological validity for the presence of a disability. This means that the C-LIM is merely one tiny part of a much larger process of nondiscriminatory assessment and that its proper use comes within a broader, comprehensive framework for evaluation that includes so much more than just examining the influence of cultural and linguistic factors on obtained test scores.

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:

1. 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. In addition, it should be apparent that the process of nondiscriminatory is a broad, multi-faceted procedure that incorporates a wide range of data and information which must be integrated in a manner that allows the preponderance of evidence to support final conclusions and inferences regarding the presence of a disability. It should be clear as well that the C-LIM is only a minor part of this broader process and that its sole function is to address the issue of test score validity. In evaluations where tests are not used, the C-LIM has no utility and cannot guide practice. But in evaluations the choose to rely on test score data, particularly in the evaluation of cognitive abilities in culturally and linguistically diverse students, the C-LIM assists practitioners in making perhaps the most important determination that may well undermine the validity of any interpretation specifying the presence of a disability—that is, to what extent to such score reflect difference vs. disorder. Because this issue is paramount to being able to use test score data in support of whatever conclusions are made, use of the C-LIM must be well understood before it can be applied as a guide for nondiscriminatory practice. The next sections offer a brief outline regarding proper application and interpretation of test scores when examined for validity within the C-LIM.

Summary

The framework for nondiscriminatory assessment of cognitive abilities outlined in the preceding section demonstrates the integrative nature of the process, particularly between the pre-referral and post-referral process. Once an assessment has been completed, it is imperative that knowledge of the individual's educational, cultural, and linguistic experiences be used to frame the patterns seen in the data, including that from test scores. With respect to standardized testing, there are several basic principles that can help create an appropriate context for expectations of performance. For example, evaluation of cultural and linguistic differences provides information that would indicate that the presence of large differences between the experiences of the student and that of monolingual, mainstream English speakers suggests a greater and more adverse effect on test performance. The more factors present in a student's background to make them more "different," the greater these differences are likely to be manifested in lower test performance. Conversely, the less opportunity a student has had for the acquisition of age-expected developmental language proficiency or acculturative knowledge, the lower the probability of the existence of a disability or disorder. It is not reasonable to have expectations of grade or age-level performance for an English learner when they have not had the same opportunity and amount of time to develop those skills or knowledge as the native English speakers to whom they are being compared.

The final step in nondiscriminatory assessment is also the most important: link results from assessment to intervention. Once assessment is completed, the student will not automatically do better in school, irrespective of whether or not a disability was identified. Therefore, the purpose in conducting an assessment should not be limited to identification efforts only; rather it should be used to inform, guide, and develop appropriate instructional interventions or

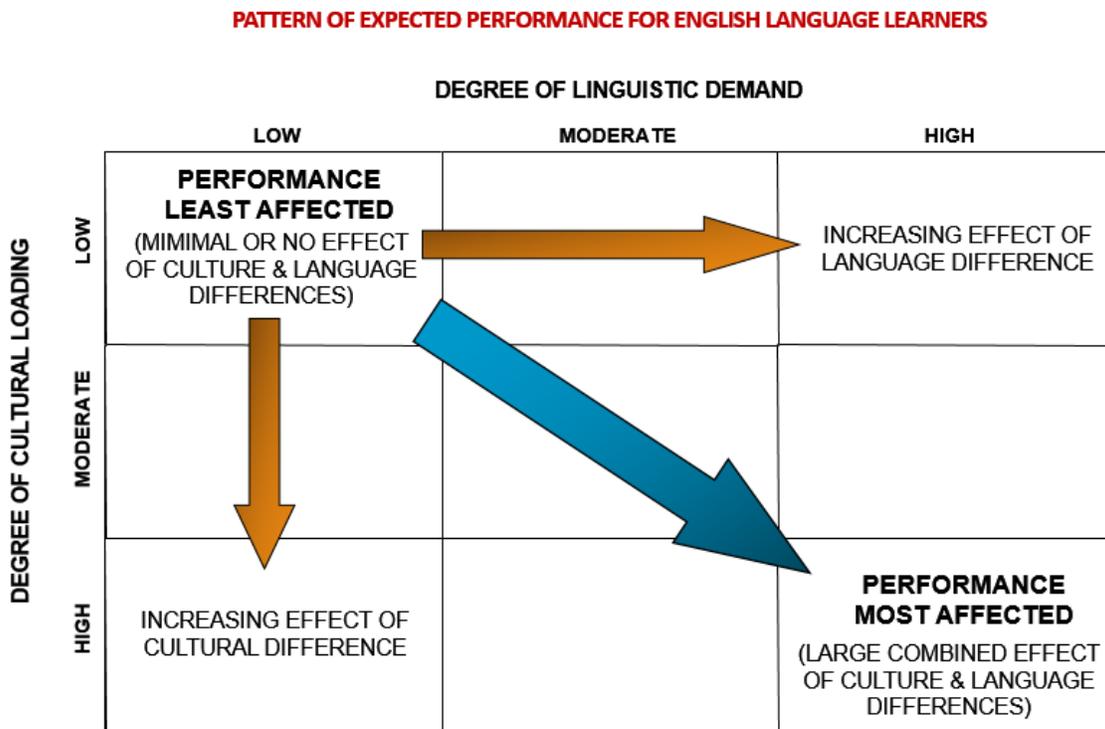
modifications, and the development of appropriate educational programs, whether under the purview of special or general education.

Instructions for C-LIM Use, Analysis, and Interpretation

Use of the C-LIM is relatively straightforward, however, the automated Excel-based versions are much easier in practice and greatly facilitate the process as well as helping avoid mathematical errors. Accordingly, the C-LIM Basic (v2.0) which is a current, semi-automated version available freely will be used for the following case study example.

In the most basic form, the C-LIM is merely a 3x3 matrix that forms nine individual cells, which can be collapsed into five separate tiers, wherein subtests from various contemporary batteries (and some legacy) are listed in one of the cells corresponding primarily to the expected level of performance as determined by subtest mean values drawn from current and historical research. The general structure of the C-LIM is provided in Figure 10.1.

Figure 10.1 General Organizational Structure of the C-LIM



Because research on how ELs perform on tests administered to them in English has long been an activity within the field of psychometrics, there is no lack of information by which to understand the patterns of performance that are typical of this group. Thus, as can be seen in Figure 10.1, the C-LIM organizes these findings by placing subtests that are least affected by cultural and linguistic factors in the upper left corner of the matrix and by placing those subtests that are most affected by cultural and linguistic factors in the bottom right corner. All other subtests

would be placed somewhere in between. This arrangement thus creates a simple expected pattern of decline in subtest scores that would begin with the highest values in the upper left descending in magnitude to the lower right. The classification and placement of subtests within the matrix is based on knowledge gleaned from over a century of research on the test performance of ELs when tested in English. For reference purposes, Table 10.4 lists an example of the mean Scaled Scores for a range of WISC subtests drawn from four different studies and provides the grand/aggregate means for each one that establish the basis of their classifications within the C-LIM as noted in the final column.

Table 10.4 Calculation of Grand Means from WISC Research Studies

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

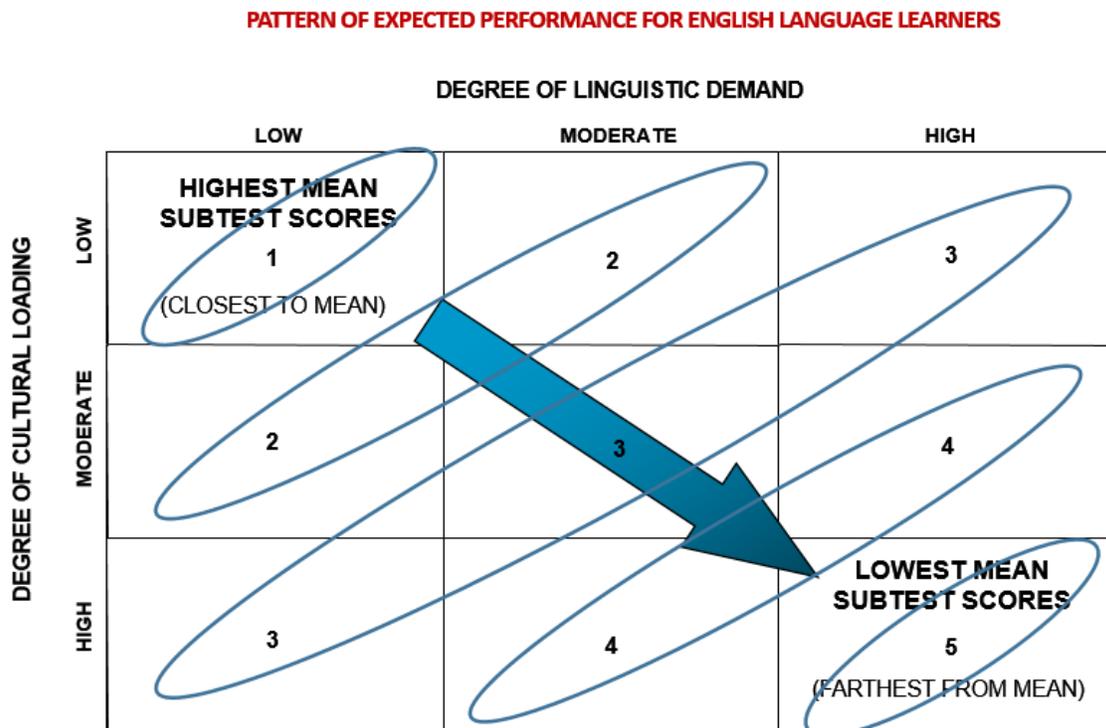
*Data for this subtest were not reported in the Nieves-Brull study.

As can be seen in Figure 10.1, regardless of the study, there is a clear pattern of performance that is related to the nature of the subtests administered to English learners. At the top of the list are subtests with very high, age-expected language and acculturative knowledge demands (e.g., Information, Vocabulary, Similarities) which contrast with the subtests at the bottom of the list which have little, if any, requirements regarding language skills or acculturative knowledge (e.g., Coding, Picture Completion, Object Assembly). Other subtests fall somewhere in between. What is remarkable about these studies is how consistent they tend to be. While there are sampling differences (e.g., the Vukovich & Figueroa sample was a subset of Mercer's sample but tests 10 years later; the Cummins' sample is from French-English immersion students in Canada) which likely account for the differences in mean scores across the same

subtest, the overall pattern where scores are lowest on language-based subtests and highest on non-language-based subtests holds true and is nearly invariant.

The purpose of the C-LIM is to take such findings (from all available studies) and collate them so as to produce the necessary mean values for classification. The classifications are organized within the C-LIM in such a manner that the single cell in the upper left corner represents Tier 1 and the single cell at the bottom right corner represents Tier 5. All other tiers (2 through 4) have more than one cell which comprises it and the middle one (Tier 3) is actually comprised of three individual cells. The subtests are also classified secondarily with respect to the degree to which each one requires age-expected levels of development regarding acculturative knowledge acquisition (cultural loading) and language proficiency (linguistic demand). These secondary classification criteria are only informational and relate to possible analyses that may examine performance differences that could be the result of task characteristics and thus may affect the position of a subtest in the matrix within a particular tier. They do not, however, determine the tier classification which is based foremost on the expected mean values as drawn from research. Figure 10.2 provides an illustration of the tiered structure of the C-LIM and identifies which specific cells belong to which particular tier.

Figure 10.2 Tiered Structure of the C-LIM

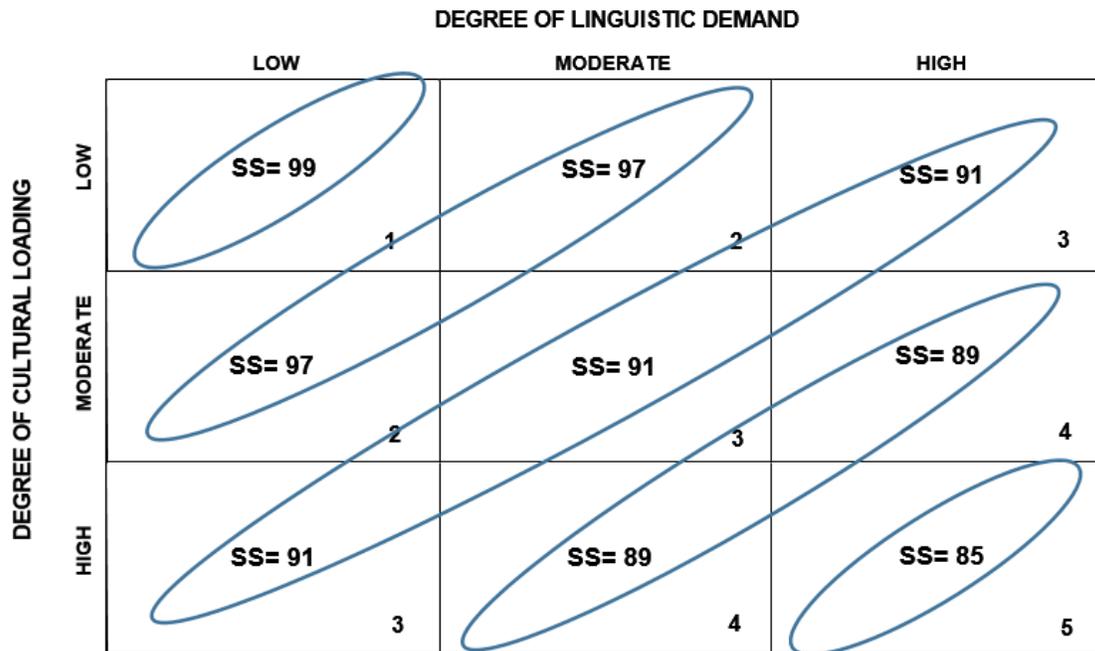


The large arrow depicted in the center of the matrix in Figure 10.2 is intended to represent the expected pattern of declining scores that have been demonstrated in research. By organizing the tests in this manner, the decline in performance, relative to the increasing demands related

to language and acculturative knowledge acquisition, can be examined from both a simple mathematical basis as well as a visual basis when graphed. Figure 10.3 provides another illustration of the matrix, with the five tiers circled around their respective cells, but with the mean values for each cell as culled from the literature.

Figure 10.3 Research-Based Expected Mean Values for C-LIM Cells/Tiers

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



There is one other consideration necessary to make the C-LIM more applicable and representative of expected “true peer” performance. Because of the lack of researchers with expertise in diversity issues or bilingual abilities, nearly all prior and current research has been conducted with highly proficient ELs so as to both eliminate the need for attention to linguistic issues as well as attempts to avoid confounds due to limited proficiency. This means that the values reported in the literature, and as shown in Figure 10.3, are those that are only appropriate and representative of the performance of individuals who can be considered to be “slightly different” than the norm samples used for comparison which were driven by monolingual English speakers. This means that individuals with proportionately less English proficiency (including those still designated by their districts as limited English proficient) will likely score more poorly. Investigations of the degree to which language proficiency attenuates test score performance for English learners are highly valuable but sorely lacking. Nevertheless, what research is available is consistent in demonstrating the drop in performance is quite minor for subtests that require little or no language/acculturative knowledge but increases linearly as tests required more age-appropriate language skills/acculturative knowledge, and that the magnitude of the difference in performance between subtests is directly related to current language proficiency (Sotelo-Dynega, Ortiz, Flanagan & Chaplin, 2013; Dynda, 2008). Therefore,

to improve the degree to which research is used to guide practice and decision-making, it is necessary to provide some distinction in the expected performance of an English learner as a function of their current language proficiency status and other related variables. Based on the aforementioned research, Table 10.5 lists the range of expected performance for each cell in the matrix and across the tiers so that users may be able to apply the most appropriate standard for evaluation of expected performance relative to language and acculturative experiences.

Table 10.5 Expected Range of Performance as a Function of Difference

	Low Language	Moderate Language	High Language
Low Culture	<ul style="list-style-type: none"> • Slightly Different: 3-5 points • Moderately Different: 5-7 points • Markedly Different: 7-10 points 	<ul style="list-style-type: none"> • Slightly Different: 5-7 points • Moderately Different: 7-10 points • Markedly Different: 10-15 points 	<ul style="list-style-type: none"> • Slightly Different: 7-10 points • Moderately Different: 10-15 points • Markedly Different: 15-20 points
Mod Culture	<ul style="list-style-type: none"> • Slightly Different: 5-7 points • Moderately Different: 7-10 points • Markedly Different: 10-15 points 	<ul style="list-style-type: none"> • Slightly Different: 7-10 points • Moderately Different: 10-15 points • Markedly Different: 15-20 points 	<ul style="list-style-type: none"> • Slightly Different: 10-15 points • Moderately Different: 15-20 points • Markedly Different: 20-25 points
High Culture	<ul style="list-style-type: none"> • Slightly Different: 7-10 points • Moderately Different: 10-15 points • Markedly Different: 15-20 points 	<ul style="list-style-type: none"> • Slightly Different: 10-15 points • Moderately Different: 15-20 points • Markedly Different: 20-25 points 	<ul style="list-style-type: none"> • Slightly Different: 15-20 points • Moderately Different: 20-25 points • Markedly Different: 25-35 points

The determination regarding the appropriate degree of difference is a decision that should be made on the basis of the data and information that has been collected in the case beginning with the pre-referral process. As noted, Step 1 of the proposed Framework for Nondiscriminatory Assessment of Cognitive Abilities, it is necessary to examine and investigate the various linguistic, educational, and experiential factors in a student’s background so as to set an appropriate context for interpretation of any subsequent test score data. From this information, an individual assessor or more appropriately a pre-referral or assessment team, can glean the necessary information by which to make such a determination. To assist in this process, the following guidelines are provided:

- Slightly Different: Includes individuals with very high levels of English language proficiency (e.g., CALP) and high acculturation, but still not entirely comparable to mainstream U.S. English speakers. Examples include individuals who are third generation in the U.S. or later, 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 commonly used 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. (the default and most common category)

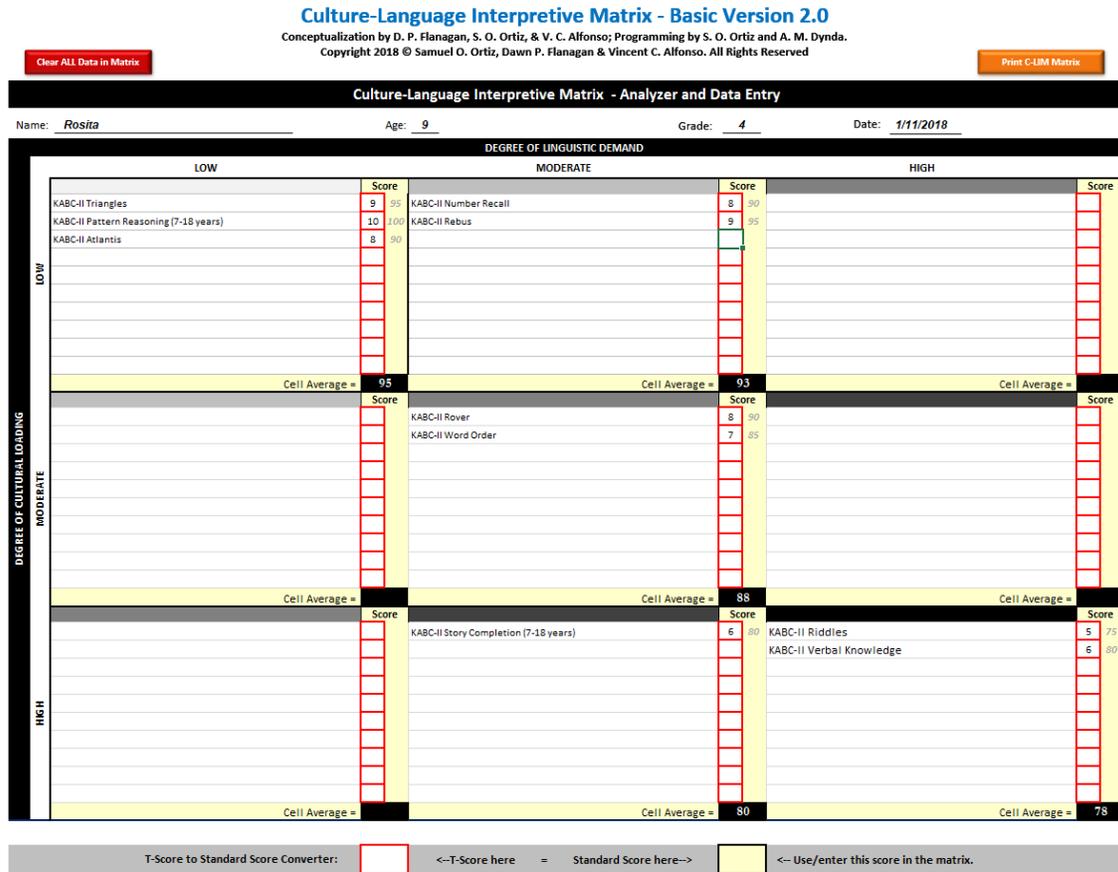
Markedly Different: Includes individuals with low to very low levels of English language proficiency (e.g., early BICS) or very limited acculturative learning experiences due to unusual influences on development. Examples include extremely low and limited parental SES and education, recently arrival 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.

Once a determination has been made regarding how "different" an individual is from the mainstream experiences and development of monolingual, mainstream English speakers in the U.S., the C-LIM can be put into use by simply entering the names and scores of any cognitive, language, and neuropsychological tests that may have been administered. In addition to the requirement that the tests be administered in English (in accordance with the proposed nondiscriminatory assessment framework) it is also important to understand that achievement tests (tests that measure specific academic skills) cannot be used in the C-LIM as there are no classifications for them. This is because the performance of an individual on a given academic skills task is dependent primarily on what grade they are in and how much formal education they have received with respect to that task/skill. Thus, the research base upon which it may be determined what is average or expected academic skill development is not the same research base upon which cognitive ability development is based. For these reasons, specific subtests that measure academic achievement or skills cannot be used in the C-LIM and are not classified within it. However, a subtest that does measure a cognitive based skill (e.g., auditory processing) but which appears on an academic achievement or mixed cognitive and achievement battery is likely classified and can be used with the C-LIM.

Although it is still possible to use the older forms for manual entry of subtest names and calculation of cell mean/aggregates, the free basic version of the C-LIM makes this unnecessary and significantly reduces the labor involved in its application. Use of the C-LIM Basic allows practitioners to merely select the name of the subtest from its cell using a simple drop down menu in the cell where the name is indicated. Each cell contains the full list of subtests that have been classified within it and a subtest appears in one and only one cell location. To determine the location of a given subtest, a complete reference list is provided in the program as well as in an appendix to this chapter, known as the Culture-Language Test Classifications (C-LTC). Enhanced versions of the C-LIM (e.g., the one contained in X-BASS) provides automatic population of subtest names from any battery either by button press or selection from a drop

down menu and further facilitates its use. An illustration of the C-LIM Basic using sample KABC-II data for Rosita are provided in Figure 10.4.

Figure 10.4 Sample KABC-II Case Data Matrix for Rosita



In this example, the KABC-II subtests that were administered to Rosita were selected from the drop down menus in the corresponding cells and their respective scores have also been entered. Note that the C-LIM permits direct entry of either Scaled Scores or Deviation IQ scores and will convert the former to the metric of the latter to enable mathematical operations. If a test such as the DAS-II, WNV, or RIAS-2 is used, a T-Score converter is provided at the bottom as it is necessary to convert T-Scores before entry into the matrix as this operation is not performed automatically in this version. As soon as the subtest scores are entered, the C-LIM will automatically calculate a mean score or aggregate for each cell. In cells with only one score, that score is used as the aggregate. This aggregate score must not be construed as representing any particular construct or latent variable. Rather, the results merely represents a simple mathematical average of subtests that share the same level of expected performance as specified by research and to some extent their shared cultural loading and linguistic demand.

Initial interpretation of the scores for Rosita in the C-LIM is accomplished by inspection of the pattern of aggregate values that were calculated and the extent to which there is a broad, overall decline in magnitude from the upper left-hand cell toward the bottom right-hand cell in a diagonal manner. 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.

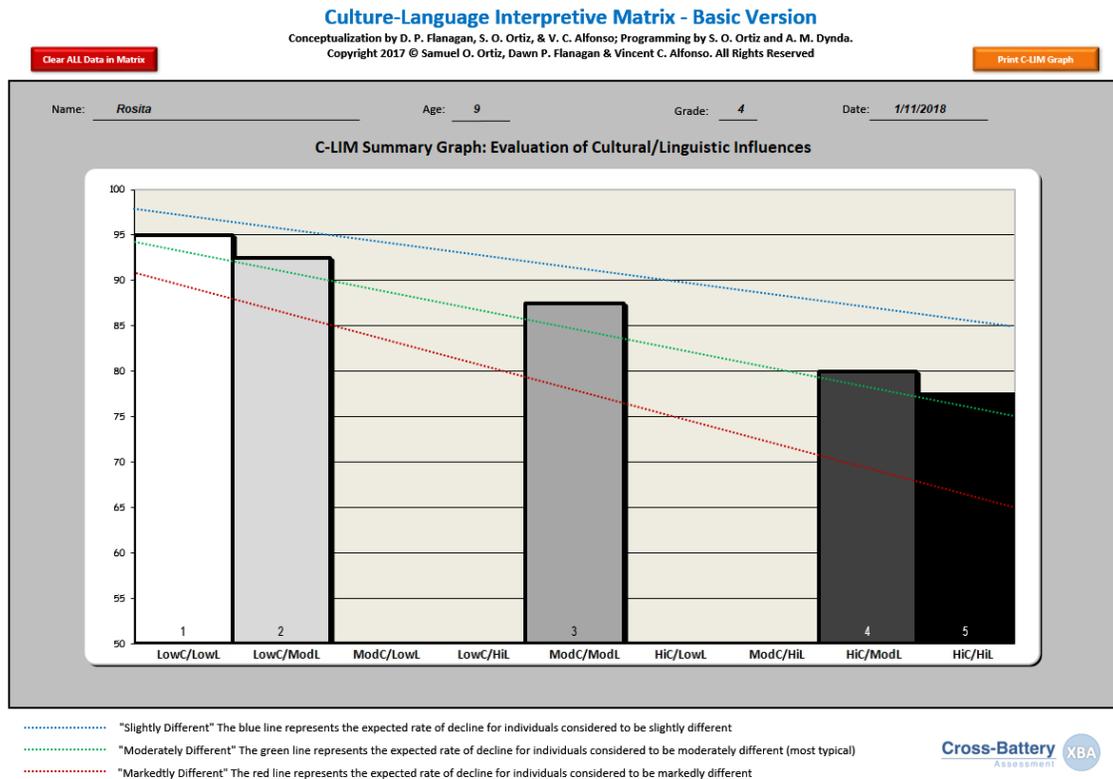
When both criteria are observed, it may be concluded that the test scores are likely to have been influenced primarily by the presence of cultural/linguistic variables and therefore are not likely to be valid and should not be interpreted. Although the C-LIM uses a two-dimensional matrix organization where one axis represents cultural loading as the other represents linguistic demand, this arrangement should not be construed as implying that these factors are uncorrelated. They are, in fact, very highly related and examination of the pattern of decline should be made primarily with respect to which the factors contribute collectively to that decline. Acculturative knowledge acquisition and English language proficiency are correlated to a substantial degree such that it is rare, albeit not impossible, that an individual will show an effect attributable only to knowledge. It is more common that language will show a more pronounced attenuation effect given its central role in all tasks involved in testing as well as the not uncommon presence of speech-language disorders. Nevertheless, the most appropriate method for inspection is to examine the aggregates in terms of decline and magnitude and as a function of the combined effect of both.

As can be seen in the analysis of scores for Rosita's KABC-II data, her highest score is in the upper left-hand corner (98) and her lowest score is also in the expected location, the bottom right-hand corner (78). The rest of her scores fall between these two extremes in a declining pattern. Thus, analysis of these results suggests that the effects of cultural loading and linguistic demand (or limited levels of acculturation and English language proficiency) were the **primary** influence on her test scores. Because these variables cannot be ruled out as the primary influences on the test results, the test results should be considered invalid indicators of actual ability. However, despite not being able to discern the individual's level of ability in any of the specific broad domains that were measured (e.g., Gc, Gf, Glr, etc.), there is a more general inference that can be made. Since Rosita's performance appears to be commensurate with the level of performance that has been established via research on ELs who are of average ability or higher and who are not disabled, it stands to reason that it is highly likely that Rosita herself has average ability or higher and does not have a disability. In other words, Rosita has performance

in the “average” range—a range that was set and established by research on other English learners with similar backgrounds as Rosita who were average and not disabled, rather than in comparison to the typical average range set by monolingual English speakers. The use of research in this case represents an evidence-based method by which to evaluate the performance by Rosita using a true peer comparison group. In so doing, it is likely that Rosita’s observed educational difficulties are not due to lack of any cognitive ability but merely reflects the normal process of second language acquisition. This conclusion is parallel to what occurs in the use of any test with native English speakers who are compared to the performance as established by the test’s norm sample. Such a comparison is not problematic for native English speakers because stratification by age and other variables effectively controls differences related to language development and opportunity for the acquisition of acculturative knowledge. This premise does not hold for English learners and so the only fair and equitable comparison that can be made must come from the application of a true peer group as constructed via existing research.

Use of the matrix and the aggregate values is difficult from a purely numerical point of view. To assist in making a clear determination regarding whether there is an overall pattern of decline and whether the scores are within the expected range, the C-LIM provides a graph to facilitate visual inspection. By plotting the nine cells linearly from left to right (low language/low culture to high language/high culture), it is possible to examine the magnitude of the cell aggregates and whether they form a declining pattern or not. By scrolling down in the C-LIM Basic, the user is presented with just such a graph that is automatically generated when subtest scores are entered and cell aggregates calculated. An example of the resulting graph for Rosita is presented in Figure 10.5

Figure 10.5 Sample KABC-II Case Data Graph for Rosita

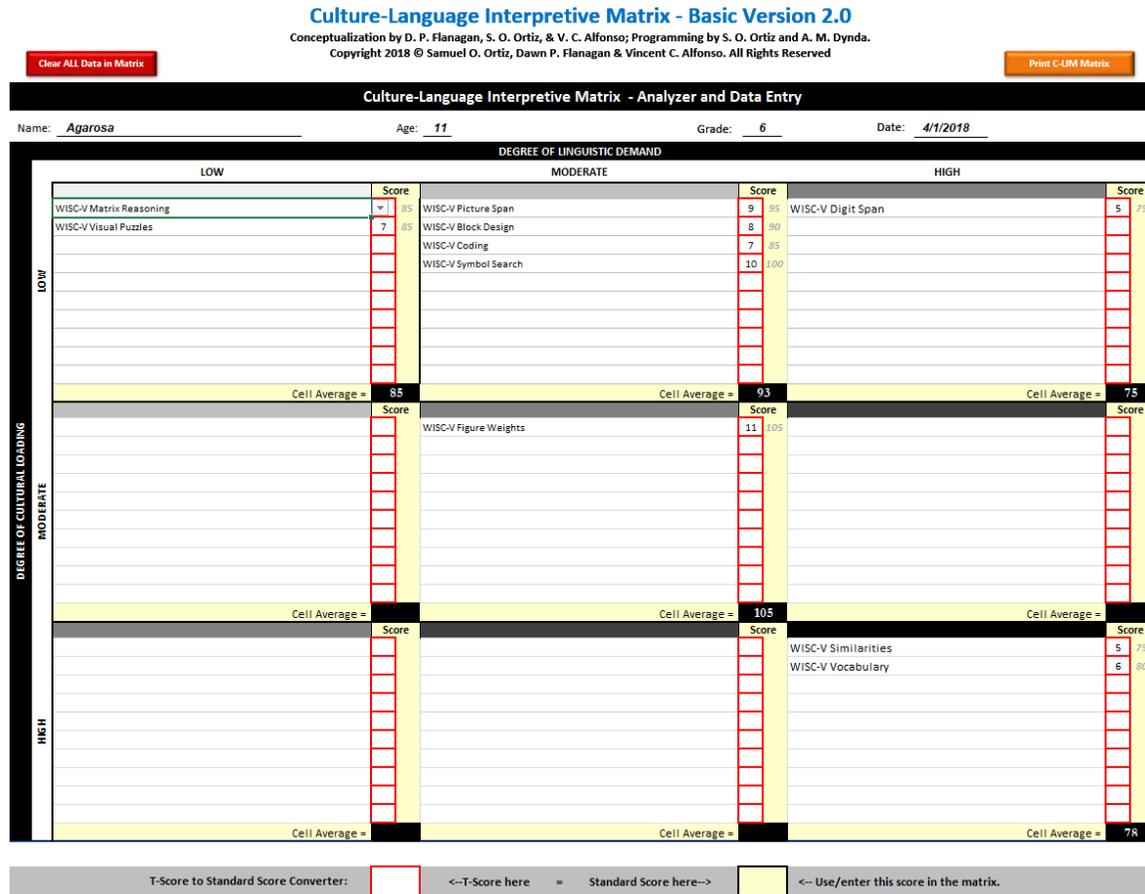


The graph illustrated in Figure 10.5 provides a visual representation of Rosita’s subtest scores that are much easier to evaluate with respect to pattern of decline and magnitude. Not only is it easier to see that the cell aggregates in this case are declining in magnitude from left to right, it is also easy to discern that the magnitude of the cell aggregates is quite consistent with the values drawn from the research and as indicated by the dotted line in the middle of the graph. This line (green in color) represents the expected pattern and magnitude of performance for English learners who have been deemed to be “moderately different” from the mainstream as is the case for Rosita. Therefore, it becomes quite clear that her scores are both declining in the manner expected and that her scores are within the “average” range expected as well. Were Rosita deemed to be “slightly different” or “markedly different,” her performance would be compared instead to the lines above (for slightly) and below (for markedly) the middle line which represent the expected values and rate of decline for individuals of such backgrounds. In this manner, the C-LIM provides flexibility in being able to more accurately assess the degree to which an individual’s scores are consistent or inconsistent with what research has established as “average” performance. Use of the graph thus provides an easier and more efficient manner for examination of test score data and, as was determined when examining the matrix, Rosita’s scores meet both basic criteria regarding the presence of an overall pattern of decline and magnitude rendering them likely invalid. But as noted, such invalidity relates only to interpretation of actual ability level, not the assessment itself, and does not preclude the conclusion that Rosita’s performance is in fact average and comparable to that of other English

learners who are of average ability and not disabled. Consequently, it can be reasoned confidently that Rosita is also likely to be of average ability as there is no evidence of deficits that might otherwise support the presence of a disability.

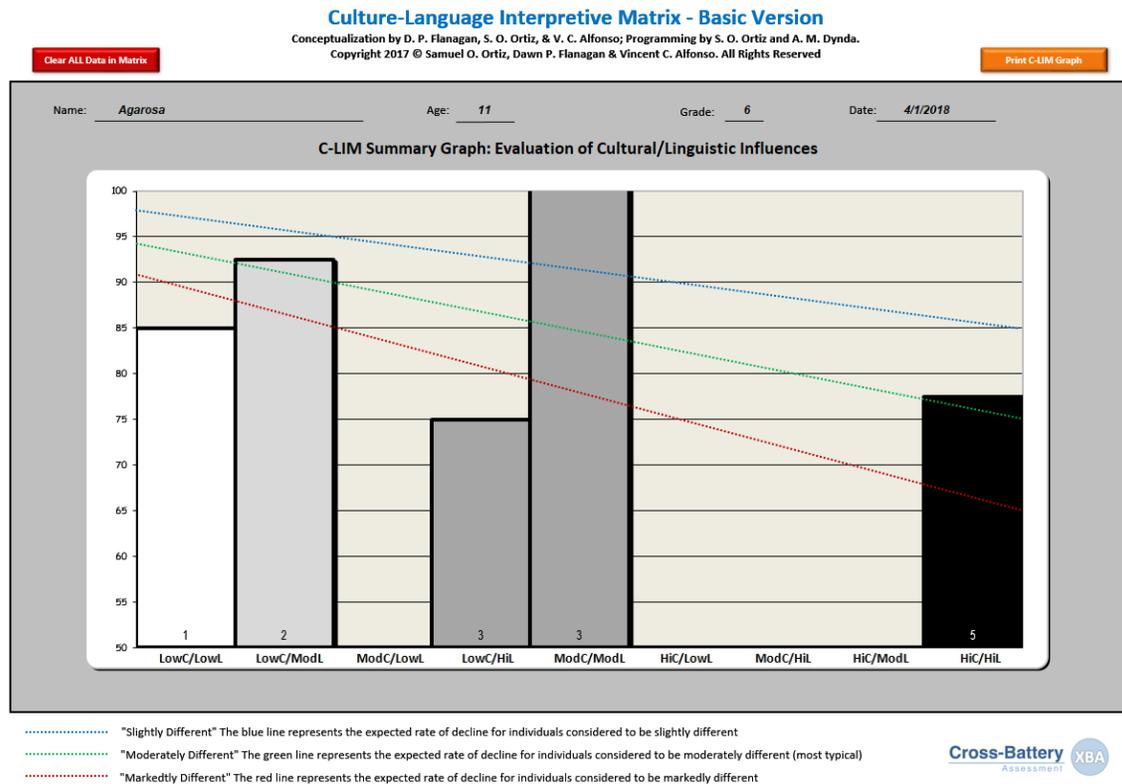
In contrast to Rosita's performance, sample WISC-V data for Agarosa illustrated in Figure 10.6 suggests a different conclusion. An examination of the pattern of results within the C-LIM for Agarosa does not appear to represent the same type of performance as was observed in Rosita's case. As in Rosita's case, initial inspection of the cell aggregate values in the matrix is readily accomplished by examining the location of the highest and lowest scores. If the pattern truly represents the primary confounding effect of culture and language, the highest score should be in the *upper left-hand corner*. In this case, it is not and instead is located in the central cell (105). Additionally, the lowest score should be in the *lower right-hand corner* but instead appears in uppermost right-hand cell (75). It is clear that whatever may have influenced the individual's performance on these tests, it was not likely to be *primarily* limited English proficiency or lack of acculturative knowledge. The results simply do not demonstrate that these factors are creating a direct and systematic influence on Agarosa's test performance. It is possible, of course, that culture and language are *contributing* factors in this case, but it is clear that they are **not primary** and that some other variable has influenced the results more than culture and language.

Figure 10.6 Sample WISC-V Case Data Matrix for Agarosa



While it is relatively easy to determine that Agarosa’s performance is not in any way commensurate with what would be expected of an average English learner with similar linguistic and cultural backgrounds, use of the graph continues to provide valuable assistance in making such determinations. When compared to the results of the graph illustrated in Rosita’s case, the resulting graph in Agarosa’s case, as illustrated in Figure 10.7 is substantially different in that it is very clear that there is no overall pattern of decline relative to cultural/linguistic factors. Indeed, her performance on some of the more heavily language-based abilities are not much different than other abilities which require virtually no language or acculturative knowledge. Likewise, some of her abilities are well below the expected range as indicated by the middle line that represents her degree of difference. In this case, both criteria for determining the impact of cultural and linguistic variables suggest that language and cultural factors cannot be viewed as being primarily responsible for the performance reflected in the cell aggregate scores. Thus, because cultural and linguistic influences can be excluded as probable confounding influences on test performance, it is reasonable to conclude that the subtest values are likely to be valid indicators and good estimates of the individual’s ability (with the exception of Gc that was discussed previously).

Figure 10.7 Sample WISC-V Case Data Graph for Agarosa



As a final consideration, it must not be assumed that establishing the validity of the test scores relative to cultural and linguistic factors does not mean that all sources of potential invalidity have been ruled out. The lack of an overall pattern of decline or aggregate scores of a lower magnitude than expected in the C-LIM can only implicate the influence of cultural and linguistic factors. Other factors that may serve to affect or attenuate test performance must also be considered as any one of them may well cause the disruption in the expected pattern or magnitude of the scores. This represents yet another reason that the proposed framework for nondiscriminatory assessment requires administration according to a tests standardization protocol—violation of standardization may be one such factor that affects performance and thereby precludes use of the C-LIM. By adhering to standardization, evaluators can more confidently rule out potential confounds related to such things as motivation, fatigue, idiosyncratic cultural response styles, as well as behavioral and attentional issues. Even simple errors in scoring, use of incorrect age norms, and the like can influence test results in an unpredictable manner that cannot be examined or evaluated with the C-LIM.

Therefore, use of the C-LIM is not in and of itself, a diagnostic tool or endeavor. Its use merely provides a basis for complying with statutory and professional regulations that require consideration of the extent to which cultural and linguistic may have affected test results and contributed to any observed educational difficulties. The C-LIM only provides a systematic method for making this determination which, while important and crucial, is not by itself

sufficient to establish the presence of a disability. What factor or variable precisely that may have been primarily responsible for the observed pattern (or lack thereof) or low score magnitude, represents a determination that must be supported with other additional sources of data. A disability may well be the reason that the pattern or magnitude of the scores in the C-LIM are the way they are but it is not the only possible reason. What the C-LIM has added to practice is only the ability to address and answer, in an evidence-based manner, the question of difference vs. disorder. Once “difference” has been ruled out (by the presence of valid scores), it is still necessary for practitioners to use all available data to draw supportable and valid conclusions regarding the presence of a disability. Of course, without the C-LIM such a determination cannot be made using test scores and may well limit the ability to draw more conclusive and defensible inferences from other data.

Additional Considerations Regarding the C-LTC and C-LIM

Although the C-LTC and C-LIM are designed as aides for engaging in evidence-based practice, the research that forms the evidence base is constantly changing and in some cases is simply lacking. Most research with English learners, as with native English speakers, is performed with the more popular tests in use. It is certain that some tests have never been used in an empirical investigation with English learners. As such, users of the C-LIM should remember that the culture-language classifications are not necessarily definitive or unchangeable. Given the gaps in research, the vast array of batteries used in evaluations, the continuous march of test revision, the limitations inherent in working with English learners, there are some classifications for which no empirical data exist to guide expected levels of performance. For example, when the WISC-IV was updated to the WISC-V, a number of years passed before any research on its use with ELs even began to appear in the literature. This gap tends to leave practitioners at a loss regarding what to make of the use of less popular or brand new tests in many of their necessary evaluations. Practitioner cannot merely decide not to test because it is an impractical and perhaps illegal solution. To assist practitioners in managing this gap, expert consensus procedures are used to provide preliminary classifications for those tests which have yet to be examined empirically. Not only does such a process represent established research methodology, it is also not nearly as difficult as it may seem. By examining the nature of a test, how it is administered, its content, age-level, ability being measured, and so forth, it is relatively easy to make some classifications, particularly those that are either very high in terms linguistic/cultural demands (as are all tests that measure Gc) or very low in this regard (as are most, but not all, nonverbal tests). In addition, new subtests often rely on tasks that are similar and nearly identical to existing tasks for which research already exists. Similar tests measuring similar abilities in similar ways should rightly lead to similar performance and thus similar mean values. Likewise, new revisions of existing tests carry over many of the same subtests as before. In both cases, prior classifications and established mean values serve as excellent guides for the classification and expected performance of new and retained subtests. And, of course, as new research is conducted, it is a simple matter to alter existing classifications should there be cause to do so. The aggregation of research provides a powerful basis for establishing the mean values

that primarily drive classification and assists in determining the extent to which new results suggest the need for modification or simply confirm prior classification according to mean values. Taken together, these procedures permit the C-LIM to provide classifications for approximately 120 different batteries/tests which covers nearly any test in current use for evaluation of cognitive abilities. The obvious impracticality inherent in testing English learners with varying levels of proficiency with every cognitive, language, and neuropsychological test available, means that some classifications are likely to remain based on expert consensus. But in all cases, use of research to guide determination of expected performance when testing ELs, as operationalized and organized within the C-LIM, will continue to represent the only practical method by which decisions regarding the issue of “difference vs. disorder” can be addressed in a practical, systematic, and empirically driven manner.

Frequently Asked Questions Regarding Intellectual Assessment and the C-LIM

1. Do I need to use “CHC Cross-Battery Assessment” in order to use the Culture-Language Test Classifications (C-LTC) and Culture-Language Interpretive Matrix (C-LIM) described in the chapter?

No. The culture-language test classifications are independent of the ability constructs a test is actually designed to measure. This is why there are not CHC designations for any of the subtests classified in the C-LIM. Rather, their organization is based on the degree to which they share the characteristics of cultural loading and linguistic demand rather than a particular cognitive ability, such as visual or auditory processing. Therefore, it does not matter which combination of tests or test battery are used; the C-LTC and C-LIM may still be employed to analyze and interpret the results.

2. Do I need to be bilingual to use the C-LIM?

No. The research on which the C-LIM is based is made up of studies that evaluate the abilities of English learners in English, not the native language. Therefore, the reason that it is recommended that evaluations begin in English is that it is something that all practitioners can do and because it is necessary if one is to employ the C-LIM to guide practice. Note, however, that depending on far the evaluation goes, if it reaches the point where it is necessary to re-evaluate areas of suspected weakness, being bilingual would be ideal and permit the best approach for gathering such data. Practitioners who are not bilingual would need to call in someone for assistance or resort to other procedures that minimize the need to be bilingual but which are less preferable to what can be accomplished by an individual who is bilingual.

3. Can I use the C-LIM with Spanish-language (or other language) tests?

Yes and no. Although the C-LIM contains classifications for the WISC Spanish and Bateria III, these tests are included primarily for experimental purposes and to drive further research.

There is simply no research beyond a single study (Esparza Brown, 2008) that provides any guidance on how bilingual students in the U.S. perform on tests given to them in their native language. Preliminary results suggest that there is still a pattern of decline and that it is due to cultural/linguistic factors and that there is an effect from language of instruction with those getting English only showing more attenuation. However, because there is insufficient research upon which to base appropriate expectations of performance, use of the classifications for the WISC Spanish and Bateria III should be considered experimental only and used for informational purposes and thus cannot be officially condoned.

4. Can I use the C-LIM with English language tests that were administered via the alternative, Spanish language instructions such as the KABC-II or DAS-II?

No. Administration of a test in an alternative language, even when specified by the publisher, is only valid if the test provides special norms for that type of administration. Since this is not the case with any of these tests, such administration violates the technical standardization protocol of the test and introduces potential error that can affect test performance in ways and to a degree that cannot be determined. The C-LIM should be used only when tests are administered in English and only when standardization is maintained despite the recommendations or instructions by test publishers who offer other language administrations. Likewise, scores from this type of administration or scores from native language tests should not be combined with scores from testing in English only. The only appropriate use of the C-LIM is with tests given according to strict protocol guidelines and when administered in English only.

5. Is it OK to combine results from speech-language testing with cognitive testing in the C-LIM to evaluate the pattern of performance?

Yes. It makes no difference what type of test is being administered, so long as it is not an achievement or academic skills test, any age-based standard score from any test can be entered into the matrix (space permitting) and combined with any other test score data to evaluate the overall pattern of decline and the magnitude of the cell aggregates.

6. When using the C-LIM, is it necessary to have data or scores in each of the nine cells?

No. There is no need to attempt to fill all nine cells of the matrix with scores. Doing so would require significant over-testing and for purposes that are not related to the referral concerns and may be construed as invasive. Rather, whatever information and scores are generated via the normal process of conducting an evaluation are sufficient for the purposes of determining the influence of cultural/linguistic factors on test performance. It should be noted, however, that if the tests used were primarily language-based, then there will be little, if any, information regarding performance when language demands are lower. Thus, not having information about performance on the other end of the matrix may not permit as definitive a conclusion regarding the influence of cultural/linguistic factors. As a general rule, more data are always better for the same reason that increase in the number of measurements increases reliability of the measurement. Thus, where few subtests are available in the matrix, it will be more difficult to

make a definitive determination, but simply wanting to fill in empty cells is not a reason to do more testing.

7. Many batteries place a premium on speed and quick responses. Are modifications in administration such as allowing more time recommended?

No. There should be no modifications or alterations to test administration, at least not during initial testing. The C-LIM is based on research that was conducted in a manner that adhered to each test's standardization protocol and where the test was administered in English. Any alteration or modification to the testing protocol will not permit valid application of the C-LIM and therefore there would be no way to determine the validity of the obtained test scores. However, after such testing has been completed and in cases where there are suspected areas of deficit that require follow up evaluation in the native language, the use of modifications or alterations (including use of a translator/interpreter) becomes a viable option as the purpose of the follow up testing is to provide evidence of cross-linguistic validity to the suspected area of weakness already measured in English. Such confirmatory evidence need not be based strictly on standard scores and the use of modifications or alterations that can produce qualitative data and observations that stand as support for the observed weaknesses in English is thus warranted and appropriate.

8. What abilities or constructs are measured by the C-LIM?

None. The C-LIM is not a measurement tool, instrument, or scale of any kind. The only measurements involved with the C-LIM are the test scores that are obtained via the normal process of test administration. Instead, the C-LIM merely organizes existing data and research in a manner that permits practitioners to systematically evaluate the extent to which cultural/linguistic factors may have affected performance. If these factors are deemed to have been the primary influence on test performance, then by definition the results are likely to be invalid and cannot be interpreted. If these factors are deemed instead to not have been the primary influence on test performance, then by definition, the results are likely to be valid and may be interpreted. In this way, the C-LIM is merely a way of facilitating the use of existing research to guide a clinical decision regarding the validity of obtained test scores. As such, it does not measure anything but rather serves as an aid for practitioners to apply research in their decision-making process. Thus, the C-LIM is a way of promoting evidence-based practice.

9. I've read some reviews of the C-LIM that are quite critical and indicate that it is not valid. Can you respond?

Yes. There are only a handful of studies that have attempted to examine the C-LIM and nearly all have had serious methodological errors. For example, the Kranzler et al., 2010 study attempted to find differences in adjacent cell mean values (about 3 points). Such a difference requires a sample size of 300, not merely the 46 that were used. Nevertheless, the performance of their sample (an older group educated outside the U.S.) was consistent with C-LIM classifications and they did find an overall pattern of decline as expected, despite the limitations of their sample

size. More recently, two studies by Styck and Watkins (2013, 2014) attempted to evaluate the C-LIM but failed to recognize that their sample was comprised of ELs who had already been identified as having a disability (97% of them). Thus, when they examined whether their score patterns were valid or invalid, they interpreted the results in reverse. Given their disability status, ELs should have displayed the valid pattern and in fact, they reported that 89.5% of the sample did just this. Combined with the 3.5% of their sample that did not have a disability (and which displayed the invalid pattern) it shows that the C-LIM classifications were consistent with 93% of all cases. While these issues are complicated, it is troublesome that such profound errors are being used to attempt to discredit the C-LIM when the C-LIM is only a reflection of existing research. For any study to indicate that the C-LIM is not a valid guide for practice, it is only necessary to demonstrate that ELs perform in the same manner and at the same magnitude as native English speakers on tests that are both language-reduced and language-embedded. Such a finding would contradict a century's worth of research and is extremely unlikely. Thus, research studies such as these tend to be more of a reflection of errors in understanding the nature of the C-LIM by individuals not well informed about its purpose, application, and use and whose conclusions are contradicted by their very own evidence. The studies by Sotelo-Dynega and colleagues (2013) and Dynda (2010) provide substantial support for the C-LIM which is often summarily overlooked but far more revealing with respect to providing evidence with which to guide evaluation and practice.

Resources

The following references may provide useful additional information and details regarding the nature of nondiscriminatory assessment. Many of these resources contain more detailed explanations of the various concepts outlined here and are in many cases the primary sources of information that were used to create this chapter.

For more in-depth information, the following publications by Dr. Ortiz are recommended:

- *Assessing Culturally and Linguistically Diverse Students: A Practical Guide* by Robert L. Rhodes, Salvador Hector O. Ortiz, Guilford Press, 2005.
- *Essentials of Cross-Battery Assessment*, 3rd edition, by Dawn P. Flanagan, Samuel O. Ortiz and Vincent C. Alfonso, John C. Wiley & Sons, 2013.

Articles

- Dynda, A. M. (2008). The relation between language proficiency and IQ test performance. Unpublished manuscript. St. John's University, NY.
- Esparza Brown, J. (2008). *The use and interpretation of the Bateria III with U.S. bilinguals*. Doctoral Dissertation, Portland State University, Portland, OR, 181 pages.

- Flanagan, D.P., Fiorello, C., & Ortiz, S. O. (2010). Enhancing practice through application of Cattell-Horn-Carroll theory and research: A “third method” approach to specific learning disability identification. *Psychology in the Schools*, 47(7), 739-760.
- Kovalski, J. F., Lichtenstein, R. Naglieri, J., Ortiz, S. O., Klotz, M. B. & Rossen, E. (2015). Current Perspectives in the Identification of specific Learning Disabilities. *Communiqué*, 44(4).
- Kranzler, J., Flores, C., & Coady, M. (2010). Examination of the Cross-Battery Approach for the Cognitive Assessment of Children and Youth From Diverse Linguistic and Cultural Backgrounds. *School Psychology Review*, 39(3), 431-446.
- Ortiz, S. O. (2017). [Evaluation of English Learners: Issues in measurement, interpretation and reporting](#). *The Score*, APA Division 5 (Quantitative and Qualitative Methods) Newsletter, January 2017. Available at <http://www.apadivisions.org/division-5/publications/score/2017/01/english-learners.aspx>
- Ortiz, S. O., Johnston, H. N., Wilcox, G. Francis, S. & Tomes, Y. I. (2014). The primacy of IQ subtest analysis to understand reading performance for culturally diverse groups. *Journal of Learning Disability*, 20(1), 45-54.
- 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*, 50(8), 781-797.
- 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.
- Styck, K. M. & Watkins, M. W. (2014). Discriminant Validity of the WISC-IV Culture-Language Interpretive Matrix. *Contemporary School Psychology*, 18, 168-188.

Books

- Cummins, J. C. (1984). *Bilingual and special education: Issues in assessment and pedagogy*. Austin, TX: Pro-Ed.
- Flanagan, D.P., Ortiz, S.O. & Alfonso, V.C. (2013). *Essentials of Cross-Battery Assessment* (3rd ed.). New York: Wiley Press.
- Rhodes, R., Ochoa, S. H. & Ortiz, S. O. (2005). *Assessment of culturally and linguistically diverse students: A practical guide*. New York: Guilford Press.
- Valdes, G. & Figueroa, R. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex Publishing.

Chapters and Other Publications

- Brown, J. E. & Ortiz, S. O. (2014). Interventions for English language learners with learning difficulties. In J. T. Mascolo, D. P. Flanagan, & V. C. Alfonso (Eds.), *Essentials of planning, selecting and tailoring interventions for unique learners* (pp. 267-313). Hoboken, NJ: Wiley & Sons, Inc.
- Flanagan, D. P., Alfonso, V. C. & Ortiz, S.O. (2012). The cross-battery assessment approach: An overview, historical perspective, and current directions. In D.P. Flanagan and P.L. Harrison (Eds.), *Contemporary intellectual assessment* (3rd ed.), (pp. 459-483). New York: Guilford Press.
- Flanagan, D.P., Alfonso, V.C., Ortiz, S. O. & Dynda, A.M. (2013). Cognitive assessment: Progress in psychometric theories of the structure of cognitive abilities, cognitive tests, and interpretive approaches to cognitive test performance. In D. Saklofske, C. R. Reynolds, and V. Schwean (Eds.), *Oxford handbook of child psychological assessment* (pp. 239-285). New York: Oxford University Press.
- NASP (2015). [Position statement: The provision of school psychological services to bilingual students](https://www.nasponline.org/x32086.xml). Retrieved from <https://www.nasponline.org/x32086.xml>
- Ortiz, S. O. (2004). Bilingual multicultural assessment with the WISC-IV. In A. Kaufman & D. P. Flanagan (Eds.), *Essentials of WISC-IV Assessment* (pp. 245-254). New York: Wiley Press.
- Ortiz, S. O. (2014). Best practices in nondiscriminatory assessment. In P. Harrison & A. Thomas (Eds.), *Best practices in school psychology VI: Foundations* (pp. 61-74). Bethesda, MD: National Association of School Psychologists.
- Ortiz, S. O. (2011). Separating cultural and linguistic difference (CLD) from Specific Learning Disability (SLD) in the evaluation of diverse students. In D. P. Flanagan and V. C. Alfonso (Eds.), *Essentials of Specific Learning Disability identification* (pp. 299-326). Hoboken, NJ: Wiley & Sons, Inc.
- Ortiz, S. O. (2018). Bilingual multicultural assessment with the WISC-IV. In D.P. Flanagan & A.S. Kaufman (Eds.), *Essentials of WISC-IV assessment* (2nd ed.) (pp. 295-309). Hoboken, NJ: John Wiley.
- Ortiz, S. O. & Gardner, K. (2017). The culturally competent school psychologist. In M. Thielking and M. T. Terjesen (Eds.), *Australian handbook of school psychology* (pp. 81-110). NY: Springer Books.
- Ortiz, S. O., Ortiz, J. A. & Devine, R. I. (2016). Use of the WJ IV with English language learners. In D. P. Flanagan & V. C. Alfonso (Eds.), *WJ IV clinical use and interpretation* (pp. 317-354). New York: Elsevier Press.
- Ortiz, S. O. & Melo, K. (2015). Evaluation of intelligence and learning disability with Hispanics. In K. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 109-134). Washington DC: APA Books.

- Ortiz, S. O., Douglas, S. & Feifer, S. G. (2013). Bilingualism and written expression: A neuropsychological perspective. In S. G. Feifer (Ed.) *The neuropsychology of written language disorders: A framework for effective interventions* (pp. 113-130). Middletown, MD: School Neuropsych Press
- Ortiz, S.O., Ochoa, H.S. & Dynda, A.M. (2012). Testing with culturally and linguistically diverse populations: Moving beyond the verbal-performance dichotomy into evidence-based practice. In D.P. Flanagan and P.L. Harrison (Eds.), *Contemporary intellectual assessment* (3rd ed.) (pp. 526-552). New York: Guilford Press.

Other Resources

- **Ortiz, S.O. (2018).** [Culture-Language Interpretive Matrix \(C-LIM\)](http://facpub.stjohns.edu/~ortiz/CLIM/). Retrieved from: <http://facpub.stjohns.edu/~ortiz/CLIM/>

A basic, Excel-based version of the C-LIM and its attendant classifications and other relevant documents are available freely for download.

- **Flanagan, D. P., Ortiz, S. O., & Alfonso, V.C. (2018).** *Cross-Battery Assessment Software System (X-BASS)*. Hoboken, NJ: John C. Wiley & Sons.

A fully automated and enhanced version of the C-LIM is contained in X-BASS is available commercially from the Wiley.com website and provides significantly more guidance and operational features.

- **Ortiz, S. (2018).** *Ortiz Picture Vocabulary Assessment Test (Ortiz PVAT)*. North Tonawanda, NY: MHS Assessments.

The Ortiz PVAT is a receptive vocabulary assessment, with dual norms, that can be used with both ELs and fluent English speakers.

Tools

- **Tool 10.1: The Culture-Language Test Classifications Full Reference List**