The Assessment of Culturally and Linguistically Diverse Populations: a fifty year dilemma:

What progress has been made, what issues remain?

WSASP
Webinar Series 2016

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

A Very Brief History of Language and Education in the U.S.

The U.S. has always been a multilingual society and the recent "English Only" movement was fostered as a result of its connection to patriotism and jingoism, particular as associated with foreign wars.

A Very Brief History of Language and Education in the U.S.

Since schools in the U.S. are public, they are governed by public sentiment, views regarding languages other than English. This view has not been kind in public schools.

A Very Brief History of Language and Education in the U.S.

In 1974, the U.S. Supreme Court in Lau v. Nichols attempted to remedy some the inequity of English-only ("sink or swim") programs for English learners. The final ruling, based on Title VI of the Civil Rights Act of 1964 that prohibited discrimination by any agency receiving federal funding, was as follows:

> "Under these state-imposed standards, there is no equality of treatment merely by providing students with the same facilities, textbooks, teachers, and curriculum; for students who do not understand English are effectively foreclosed from any meaningful education. Basic English skills are at the very core of what these public schools teach. Imposition of a requirement that, before a child can effectively participate in the educational program, he must already have acquired those basic skills is to make a mockery of public education. We know that those who do not understand English are certain to find their classroom experiences wholly incomprehensible and in no way meaningful (414 U.S. 563).

A Very Brief History of Language and Education in the U.S.

States have enacted laws that have effectively mandated "English Only" in the schools. Recently, California, Massachusetts, and Arizona did so, but the practice existed long before in other places like Nebraska.

In 1919, Robert Meyer was charged with violating a Nebraska State law that mandated English-only instruction in all public and private schools because he attempted to teach a bible story to a 10-year-old student using German.

The State Supreme Court argued: "the Legislature had seen the baneful effects of permitting foreigners, who had taken residence in this country, to rear and educate their children in the language of their native land. The result of that condition was found to be inimical to our own safety." (262 U.S. 390).

Fortunately, the U.S. Supreme Court overturned the State decision.

A Very Brief History of Language and Education in the U.S.

As a result of Lau v. Nichols decision, States are required to identify students who lack such proficiency in English and to provide them a "special linguistic program" presumably designed to help them attain the necessary proficiency. But the court's understanding of such proficiency was sorely lacking:

> "Against the possibility that the Court's judgment may be interpreted too broadly, I stress the fact that the children with whom we are concerned here number about 1,800. This is a very substantial group that is being deprived of any meaningful schooling because the children cannot understand the language of the classroom. *We may only guess as to why they have had no exposure to English in their preschool years. Earlier generations of American ethnic groups have overcome the language barrier by earnest parental endeavor or by the hard fact of being pushed out of the family or community nest and into the realities of broader experience* (414 U.S. 563; emphasis added).

Really?

A Very Brief History of Language and Education in the U.S.

The reality of the educational success of immigrants in the U.S. is not one characterized by simple "hard work" and instant "assimilation." Rather, success in school has been the result of differences in wealth/privilege and prior education. Possession of these assets ensured success. Lack of them did not.

1908: - 54% of New York's native-born 8th graders went on to 9th grade compared to 34% of foreign born

- 80% of urban, native-born, white 7th graders graduated but only 58% of Italian children did

1910: - There were 191,000 Jewish children in New York schools, but only 6,000 were in high school, and the overwhelming majority dropped out.

1921: - Half of all "learning-disabled" children in New York "special-education" classes were Italian

1931: - Only 11% of Italians graduated high school compared with 40% overall

1957 vs. 1965 Cuban immigrants: - The wealthy and educated vs. the poor and institutionalized

1975 vs. 1979 Vietnamese immigrants: - The politically connected vs. the "boat people"

A Very Brief History of Language and Education in the U.S.

Immigrant achievement, even with continued application of "English Only" policies has increased over the years.

1972 to 1995:

- Latino high school completion crept up from 66% to 70%
- 54% of Latino graduates now enroll in college, up from 45% (it's 64% for non-Latino whites)
- Latino high school graduates who complete college rose from 11% to 16% (for non-Latino whites it's 34%)
- Graduation rate for Mexico-born youths, age 15-17 years, is 74%
- More than 70% of Latino immigrants who came here before their sophomore year in high school go on to graduate

The reason for increasing immigrant educational achievement is the same as it was for all previous, low SES immigrant groups: greater cultural assimilation and English language exposure.

A Very Brief History of Language and Education in the U.S.

Type	Stage	Language Use
		FIRST GENERATON – FOREIGN BORN
A	Newly Arrived	Understands little English. Learns a few words and phrases.
Ab	After several years of residence – Type 1	Understands enough English to take care of essential everyday needs. Speaks enough English to make self understood.
Ab	Type 2	Is able to function capably in the work domain where English is required. May still experience frustration in expressing self fully in English. Uses immigrant language in all other contexts where English is not needed.
		SECOND GENERATION – U.S. BORN
Ab	Preschool Age	Acquires immigrant language first. May be spoken to in English by relatives or friends. Will normally be exposed to English-language TV.
Ab	School Age	Acquires English. Uses it increasingly to talk to peers and siblings. Views English-language TV extensively. May be literate only in English if schooled exclusively in this language.
AB	Adulthood – Type 1	At work (in the community) uses language to suit proficiency of other speakers. Senses greater functional ease in his first language in spite of frequent use of second.
A**B**	Adulthood – Type 2	Uses English for most everyday activities. Uses immigrant language to interact with parents or others who do not speak English. Is aware of vocabulary gaps in his first language.
		THIRD GENERATION – U.S. BORN
A**B**	Preschool Age	Acquires both English and immigrant language simultaneously. Hears both in the home although English tends to predominate.
a**B**	School Age	Uses English almost exclusively. Is aware of limitation sin the immigrant language. Uses it only when forced to do so by circumstances. Is literate only in English.
a**B**	Adulthood	Uses English almost exclusively. Has few opportunities for speaking immigrant language. Retains good receptive competence in this language.
		FOURTH GENERATION – U.S. BORN
Ba	Preschool Age	Is spoken to only in English. May hear immigrant language spoken by grandparents and other relatives. Is not expected to understand immigrant language.
Ba	School Age	Uses English exclusively. May have picked up some of the immigrant language from peers. Has limited receptive competence in this language.
B	Adulthood	Is almost totally English monolingual. May retain some receptive competence in some domains.

Source: Adapted from Valdés, G. & Figueroa, R. A. (1994), *Bilingualism and Testing: A special case of bias* (p. 16).

Understanding Language, Education, and Assessment

So what does all of this mean for school psychologists interested in the assessment of English language learners?

The validity of norm-referenced, individually administered, standardized tests is based on certain assumptions. According to Salvia & Yssledyke (1991):

"When we test students using a standardized device and compare them to a set of norms to gain an index of their relative standing, we assume that the students we test are similar to those on whom the test was standardized; that is, we assume their acculturation [and linguistic history] is comparable, but not necessarily identical, to that of the students who made up the normative sample for the test. When a child's general background experiences differ from those of the children on whom a test was standardized, then the use of the norms of that test as an index for evaluating that child's current performance or for predicting future performances may be inappropriate" (p. 18).

Understanding Language, Education, and Assessment

When do the background experiences of ELLs become comparable to that of native English speakers who comprise the vast majority of the norm sample on which the test was based?

Never. The issue is not merely one based on a specific level of language proficiency, in part because language proficiency is not a static ability but rather increases with education. Salvia and Ysseldyke further assert:

> *"When we say that a child's acculturation differs from that of the group used as a norm, we are saying that the experiential background differs, not simply that the child is of different ethnic origin, for example, from the children on whom the test was standardized" (p. 18).*

Once an ELL, always an ELL. The difference in linguistic and acculturative learning experiences between those who are exposed to only one language vs. those exposed to two (or more) can never be made "equivalent" in any real sense.

Understanding Language, Education, and Assessment

Ok, so then what do we do, I hear you cry! Is there no hope for evaluating ELLs in a fair and equitable manner?

The good news is that, yes, there is a way and it is based on simply understanding that expectations of performance must be based on the degree to which the individual being tested differs in terms of these developmental experiences as compared to the normative standard of the test being administered.

This means that the validity of any set of test scores and the degree to which they accurately reflect the individual's true ability must necessarily be based on a standard that is appropriate to the individual's development rather than one that is based strictly on the norm sample's development.

General Nondiscriminatory Assessment Processes and Procedures

I. Assess for the purpose of intervention

II. Assess initially with authentic and alternative procedures

III. Assess and evaluate the learning ecology

IV. Assess and evaluate language proficiency

V. Assess and evaluate opportunity for learning

VI. Assess and evaluate relevant cultural and linguistic factors

VII. Evaluate, revise, and re-test hypotheses

VIII. Determine the need for and language(s) of formal assessment

IX. Reduce potential bias in traditional assessment practices

X. Support conclusions via data convergence and multiple indicators

Addresses concerns regarding fairness and equity in the assessment process

Addresses possible bias in use of test scores

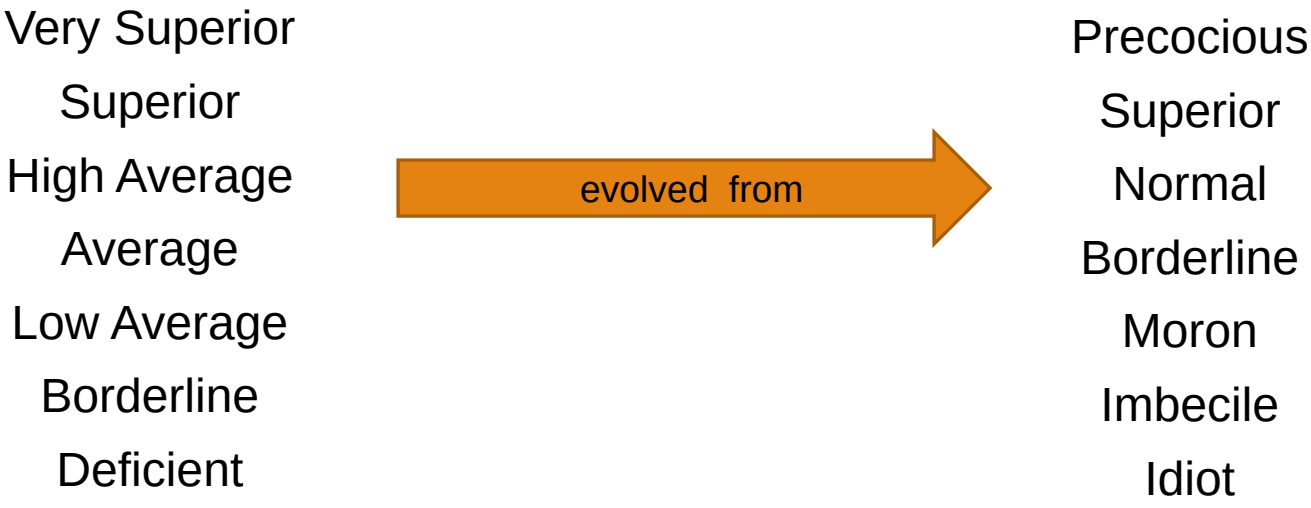
Pre-referral procedures (I. - VIII.)
Post-referral procedures (IX. - X.)

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

It was believed that:

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

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



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

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

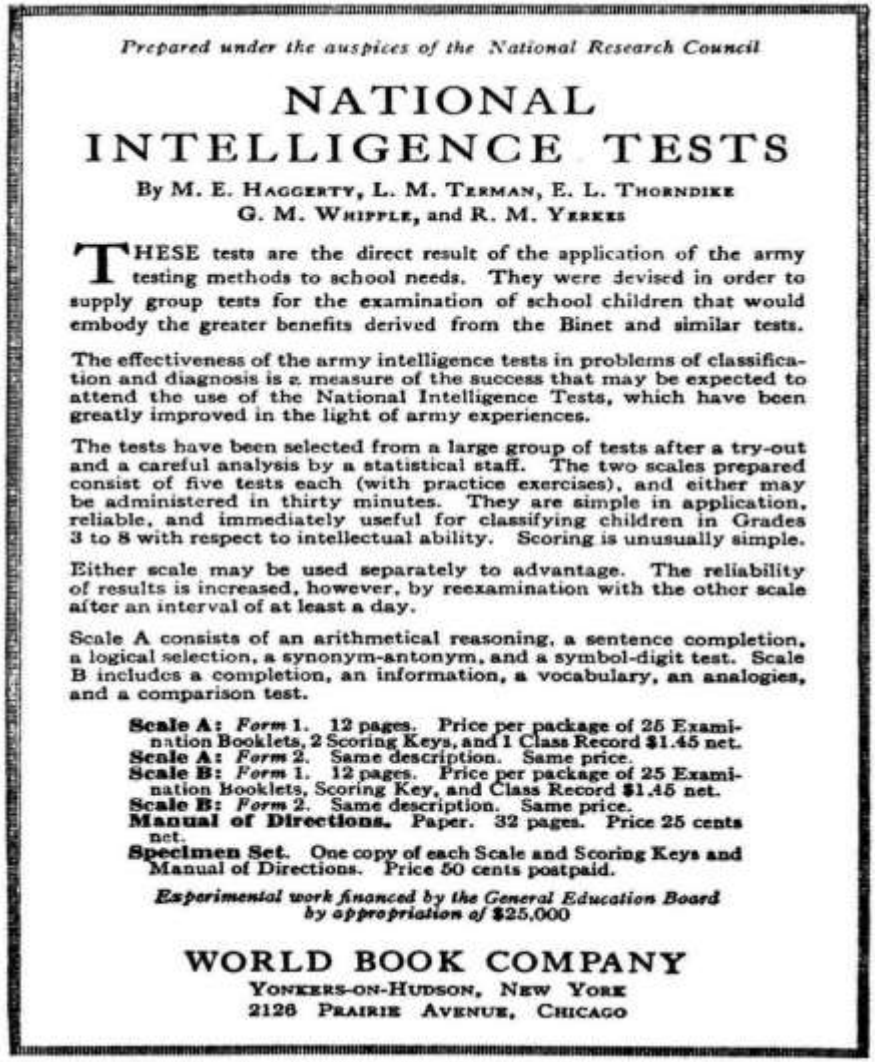
- *The testing of newly arrived immigrants at Ellis Island*

Lewis Terman and the Stanford-Binet

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

Robert Yerkes and mass mental testing

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



Prepared under the auspices of the National Research Council

NATIONAL INTELLIGENCE TESTS

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

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

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

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

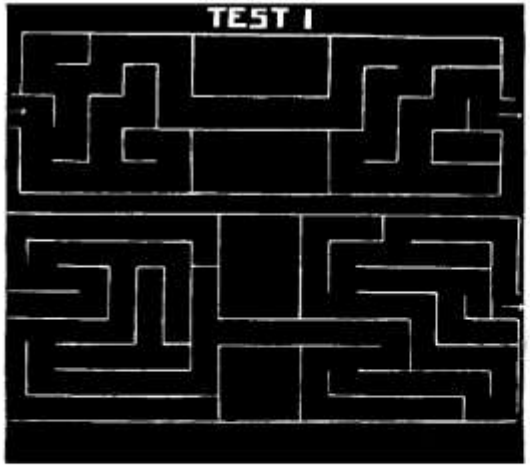
Either scale may be used separately to advantage. The reliability of results is increased, however, by reexamination with the other scale after an interval of at least a day.

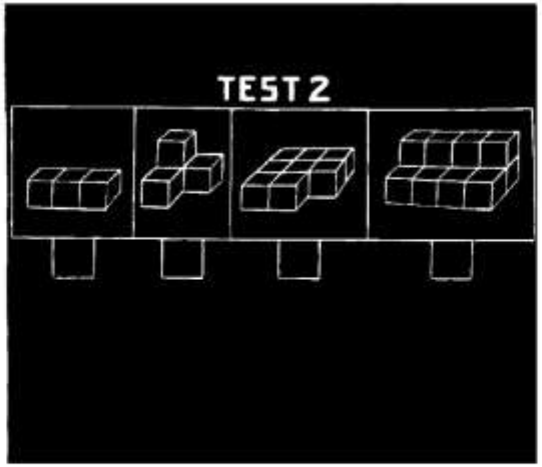
Scale A consists of an arithmetical reasoning, a sentence completion, a logical selection, a synonym-antonym, and a symbol-digit test. Scale B includes a completion, an information, a vocabulary, an analogies, and a comparison test.

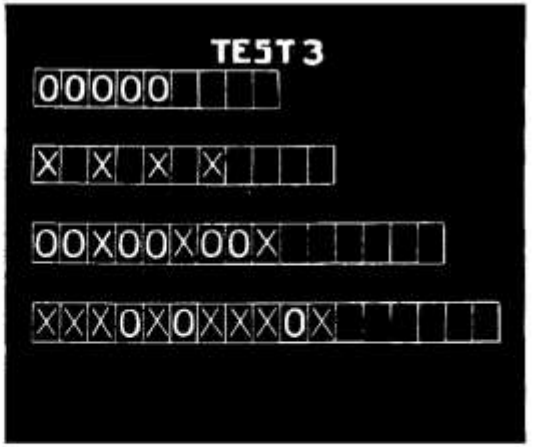
Scale A: *Form* 1. 12 pages. Price per package of 25 Examination Booklets, 2 Scoring Keys, and 1 Class Record $1.45 net.
Scale A: *Form* 2. Same description. Same price.
Scale B: *Form* 1. 12 pages. Price per package of 25 Examination Booklets, Scoring Key, and Class Record $1.45 net.
Scale B: *Form*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

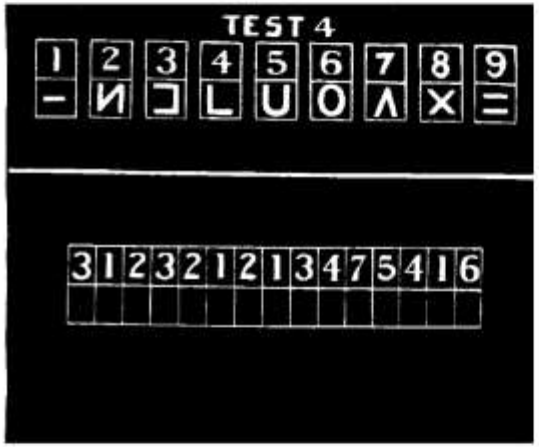
Experimental work financed by the General Education Board by appropriation of $25,000

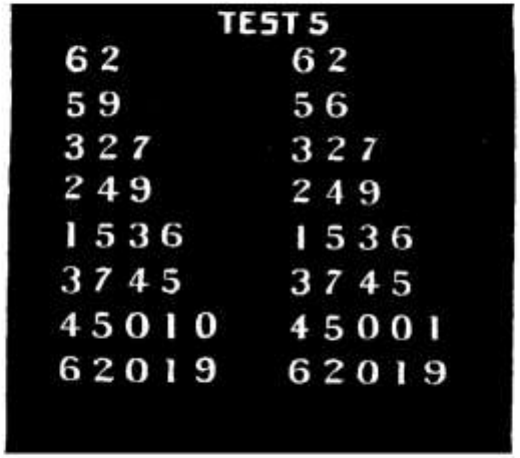
WORLD BOOK COMPANY
Yonkers-on-Hudson, New York
2126 Prairie Avenue, Chicago

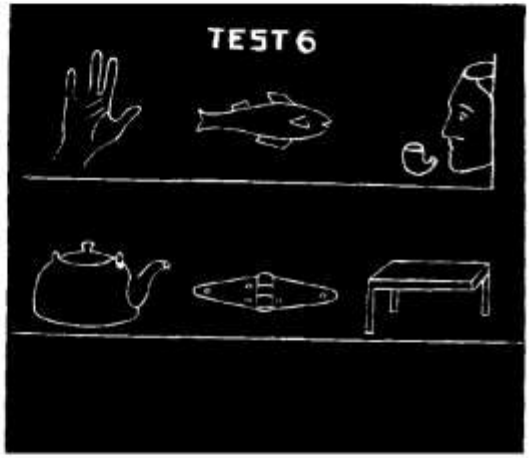


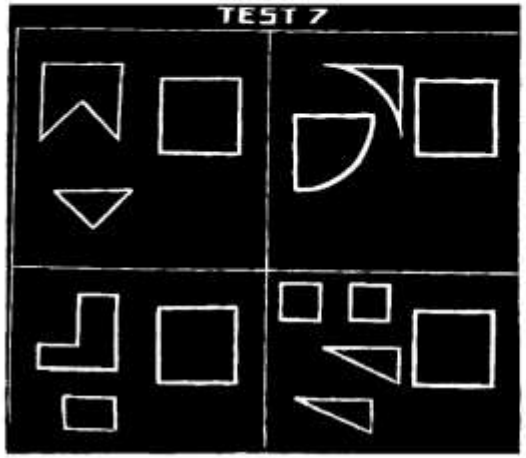




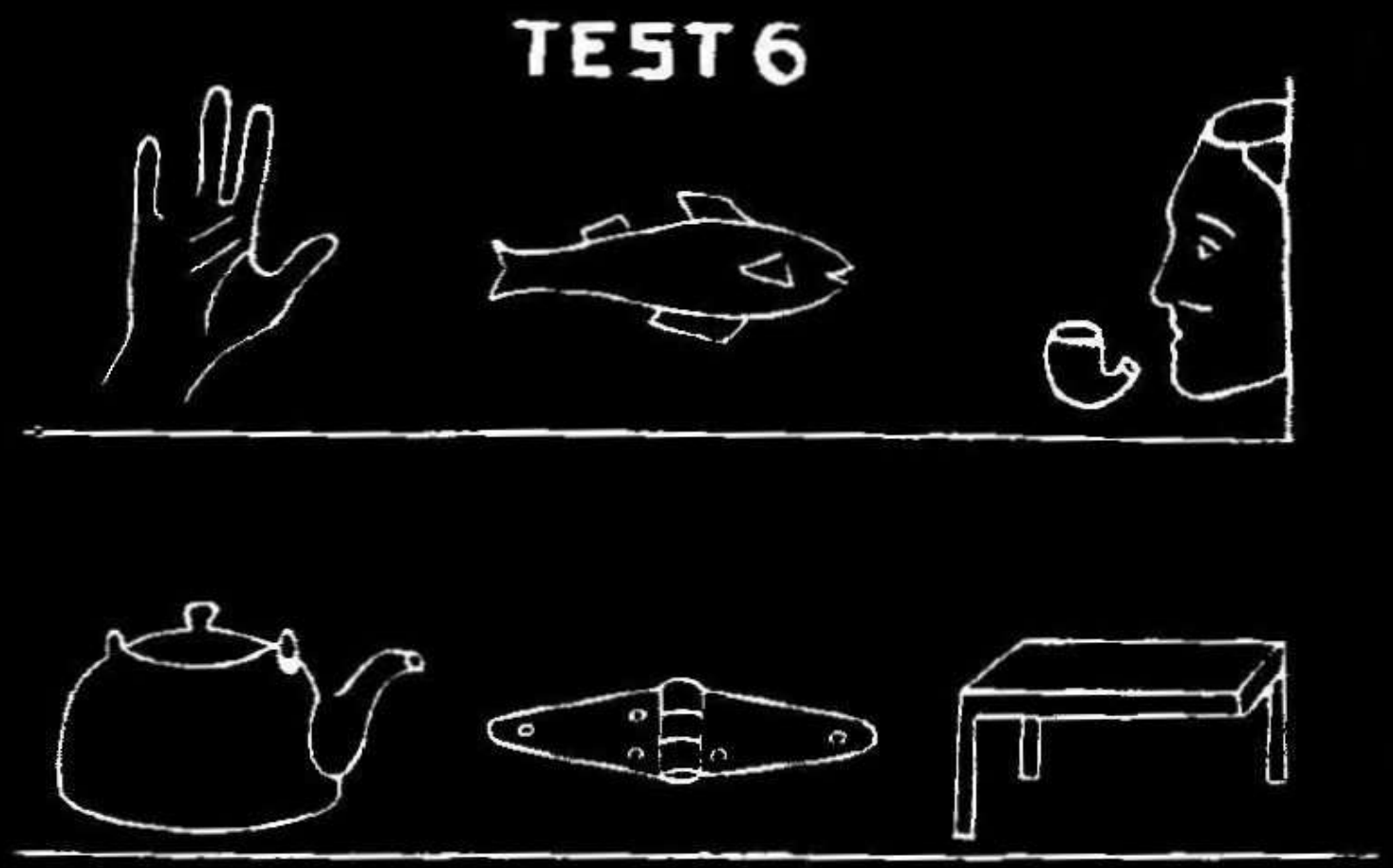






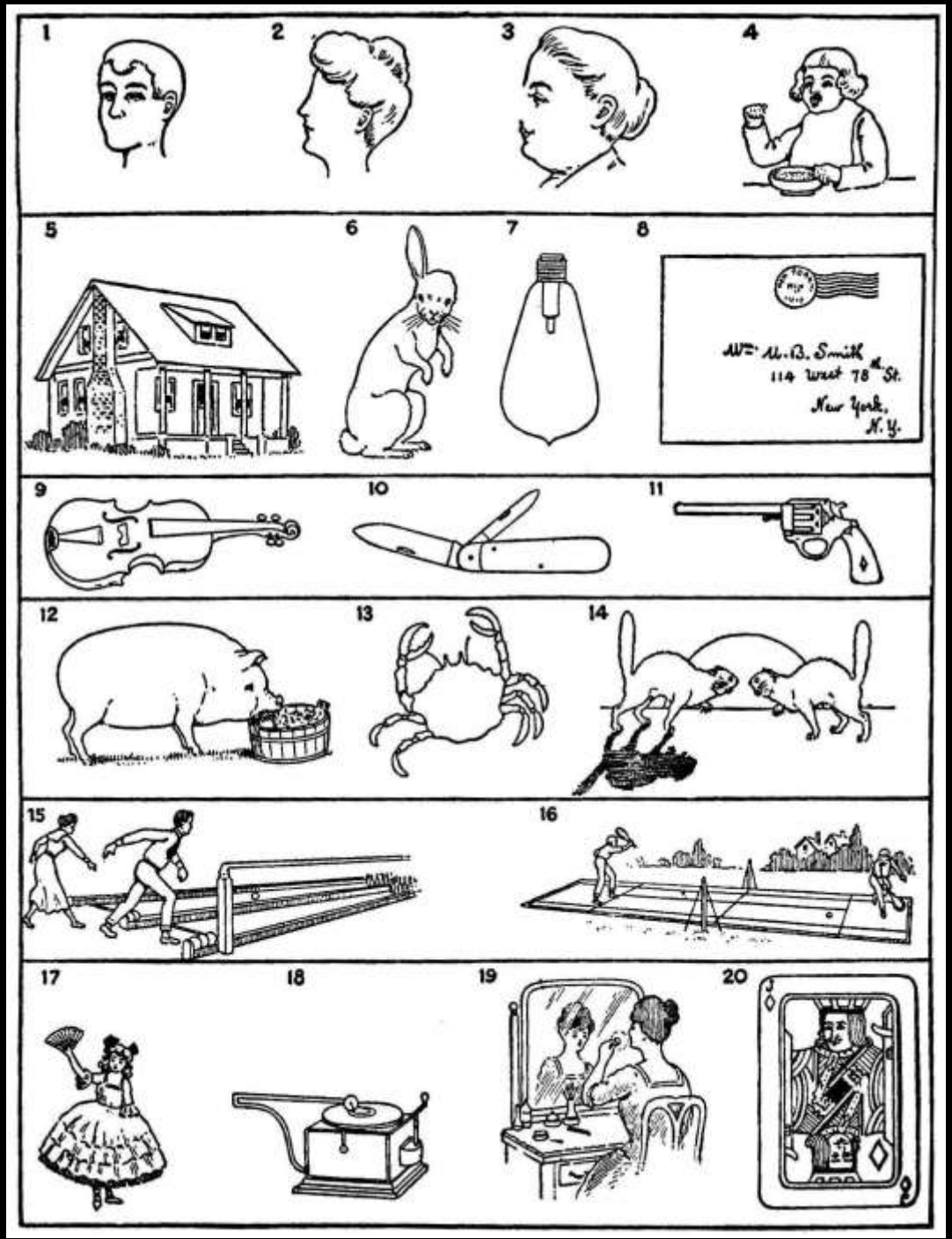


The blackboard demonstrations for seven parts of the Beta Test. From Yerkes, 1921.



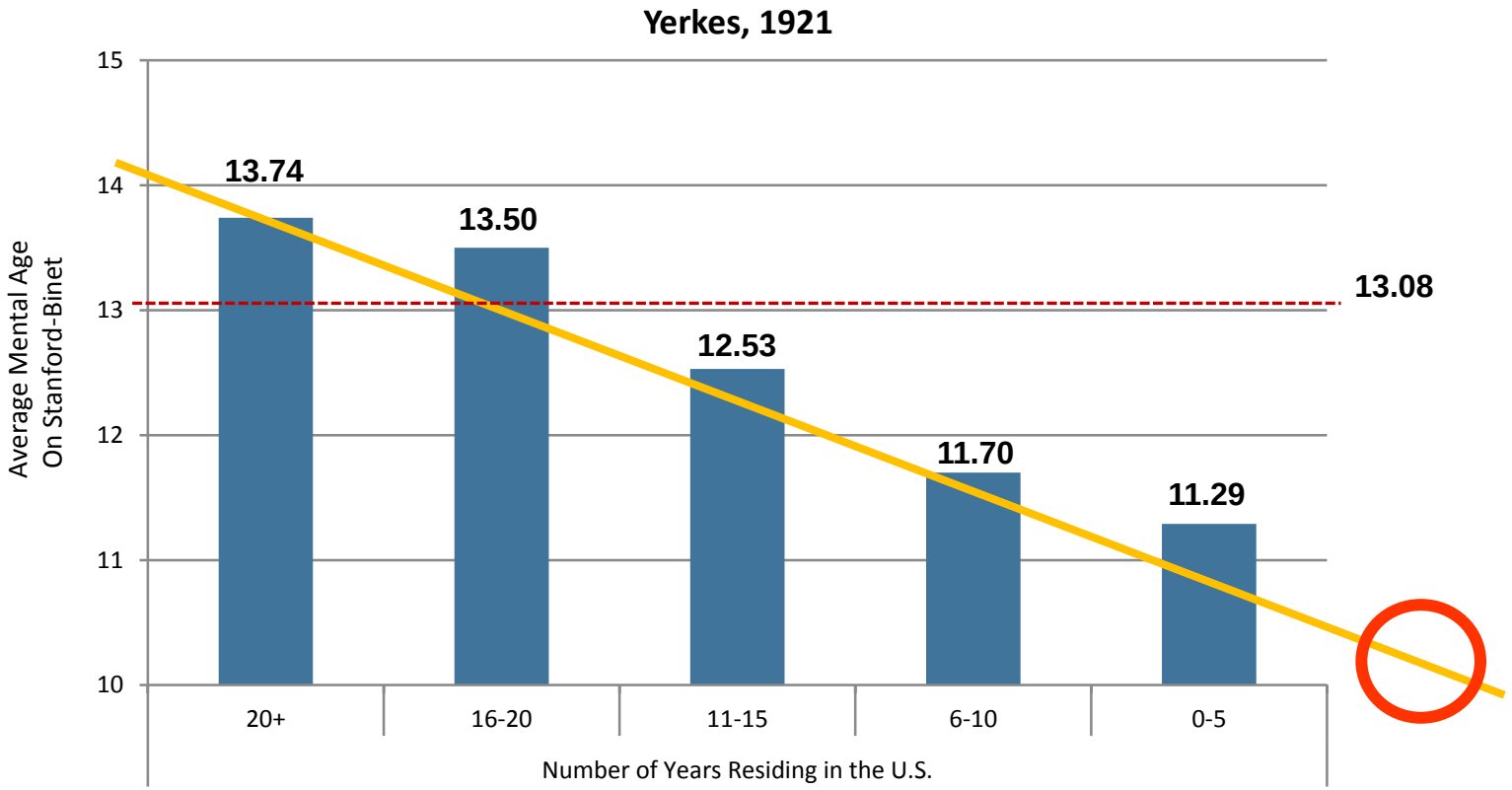
Instructional Items from Test 6 of the Army Beta Test.

Part six of examination Beta for testing innate intelligence.



A Very Brief History of Testing of English Learners in the U.S.

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



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

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

A Very Brief History of Testing of English Learners in the U.S.

- *Interpretation: New immigrants are inferior*

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

Brigham, 1923

Summary of Research on the Test Performance of English Language Learners

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

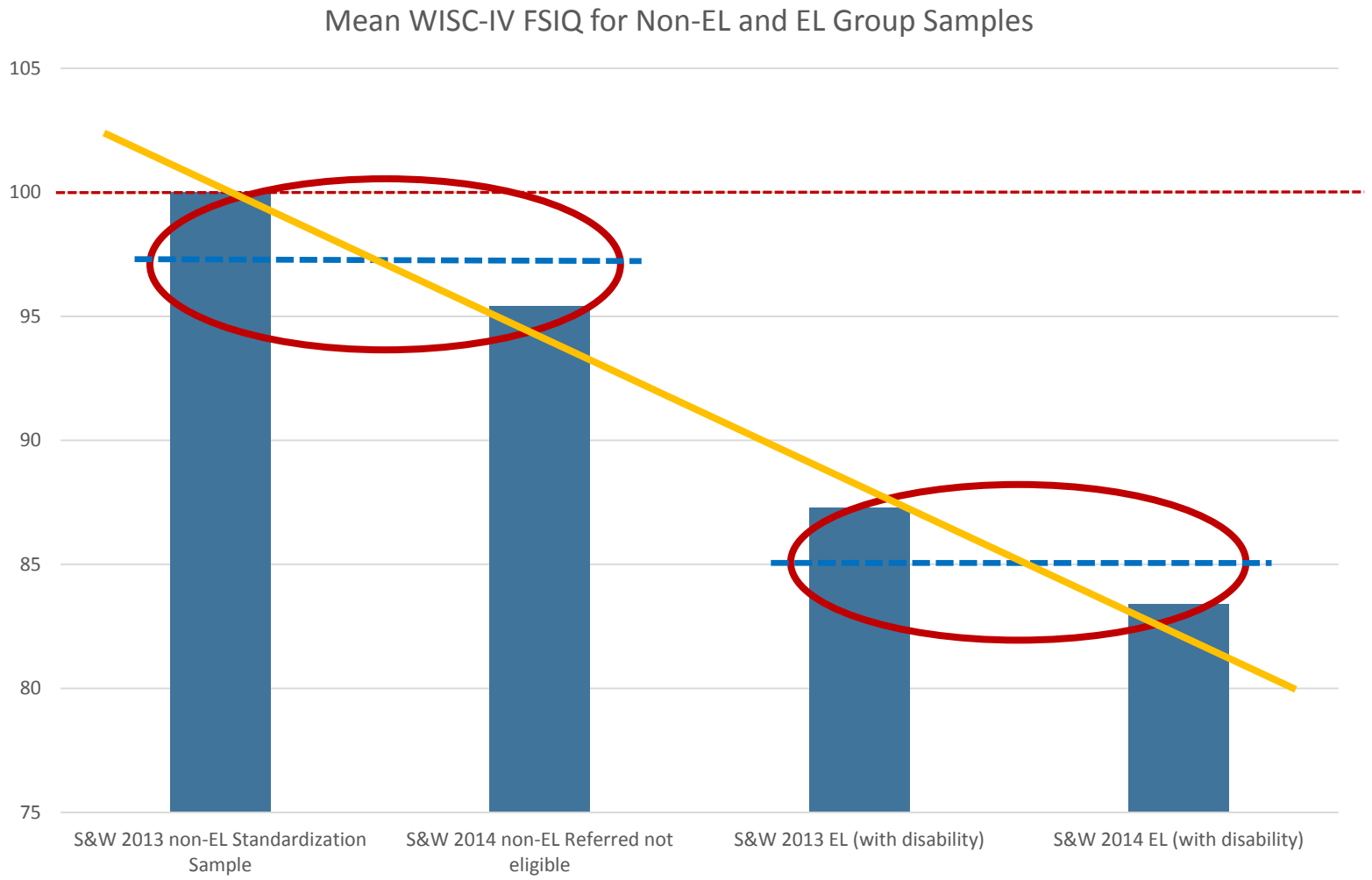
1. *Native English speakers perform better than English learners at the broad ability level (e.g., FSIQ) on standardized, norm-referenced tests of intelligence and general cognitive ability.*

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

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

Research Foundations for ELL Evaluation

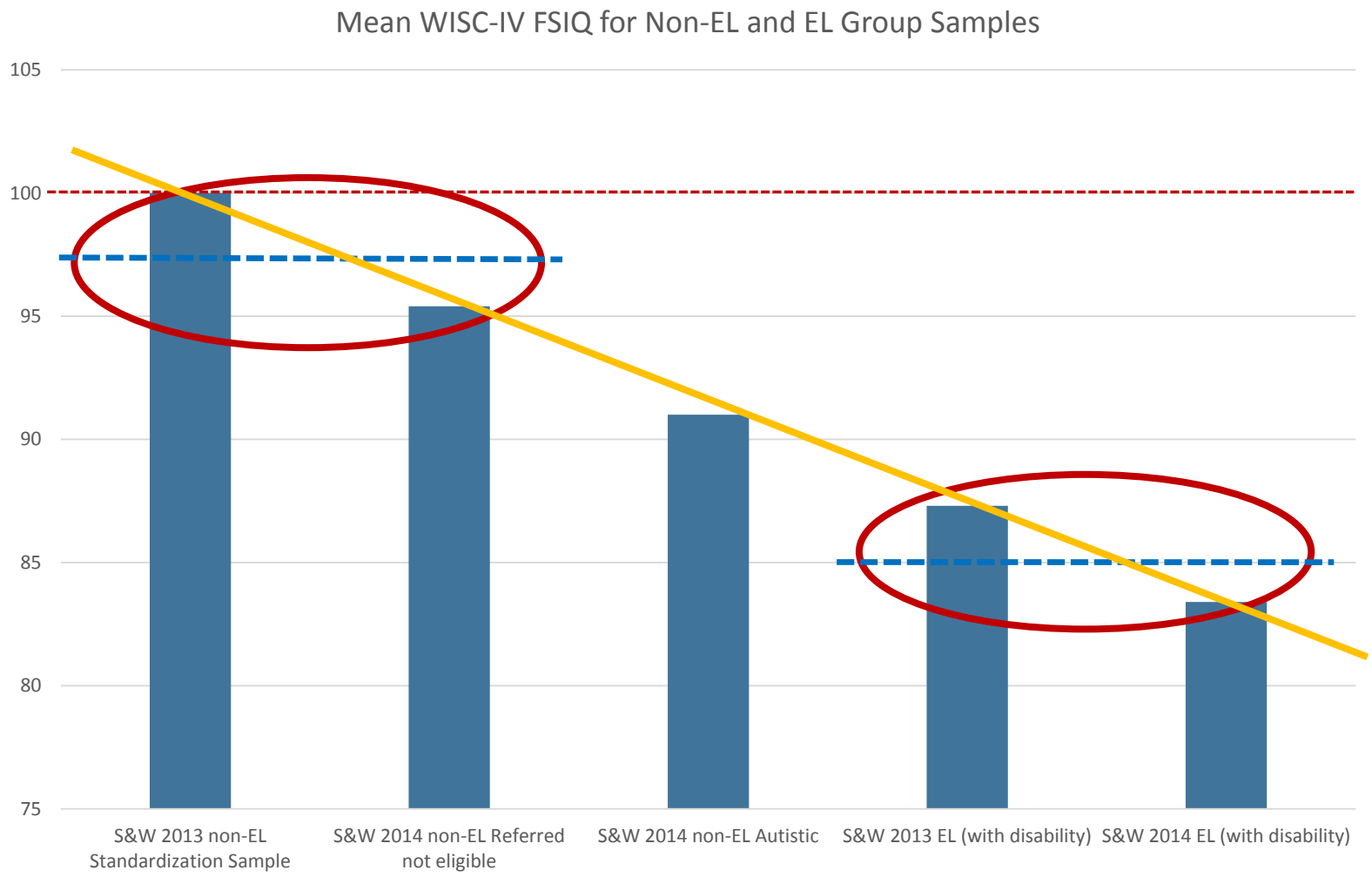
Principle 1: ELLs and non-ELL's perform differently at the broad ability level



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

Research Foundations for ELL Evaluation

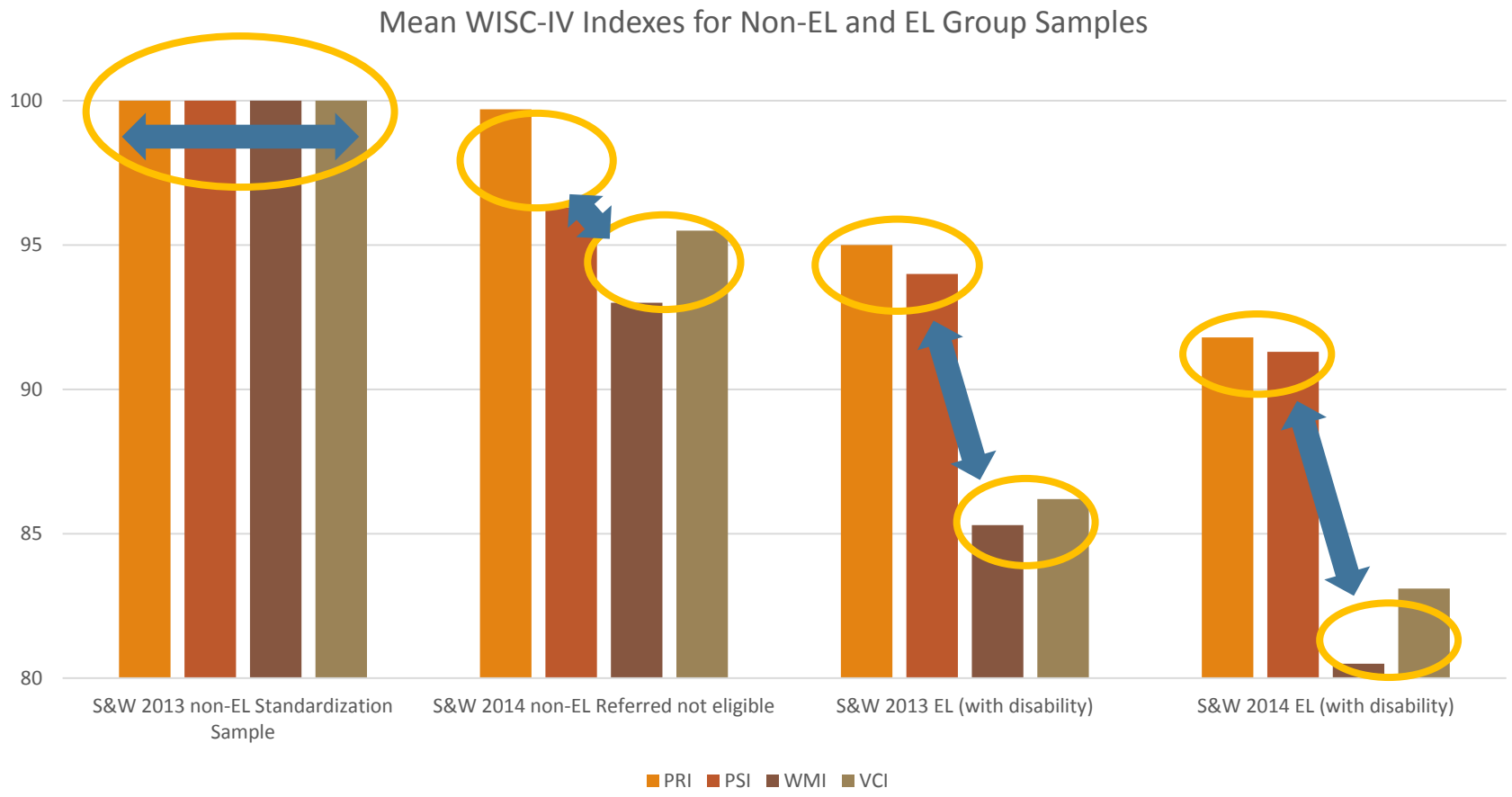
Principle 1: ELLs and non-ELL's perform differently at the broad ability level



Soures: 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. and Styck, K. M. & Watkins, M. W. (2014). Discriminant Validity of the WISC-IV Culture-Language Interpretive Matrix. Contemporary School Psychology , 18, 168-188.

Research Foundations for ELL Evaluation

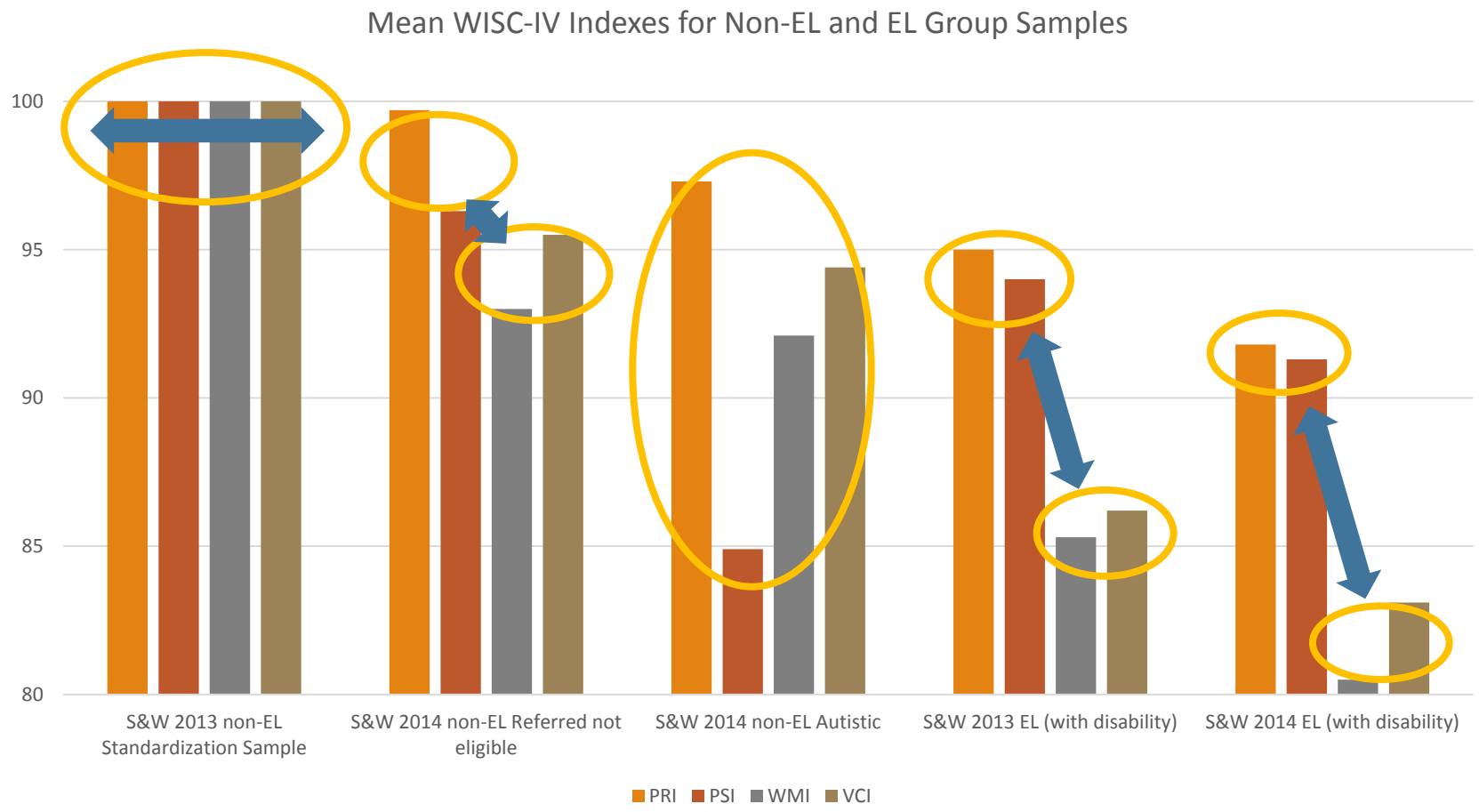
Principle 2: ELLs perform better on nonverbal tests than verbal tests



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

Research Foundations for ELL Evaluation

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



Soures: 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. and Styck, K. M. & Watkins, M. W. (2014). Discriminant Validity of the WISC-IV Culture-Language Interpretive Matrix. Contemporary School Psychology , 18, 168-188.

Research Foundations for ELL Evaluation

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

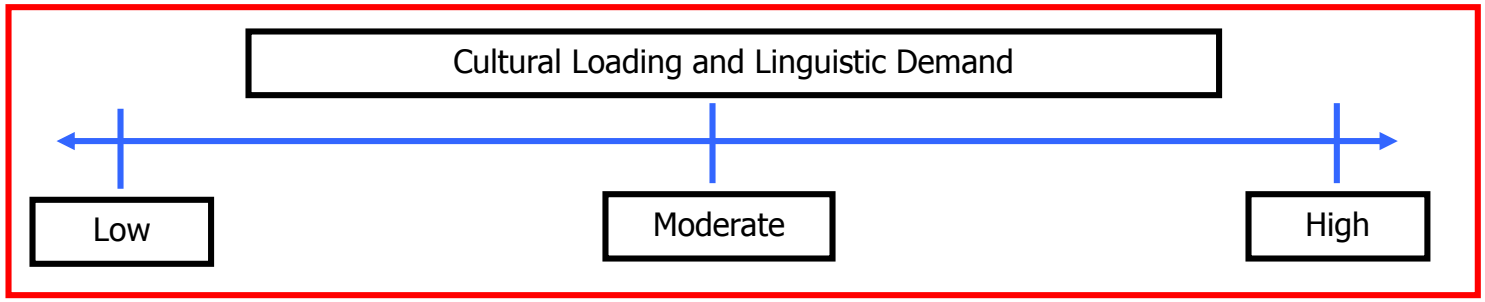
Instead, it appears that test performance of ELLs is not a dichotomy but rather a continuum formed by a linear, not dichotomous, attenuation of performance.

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

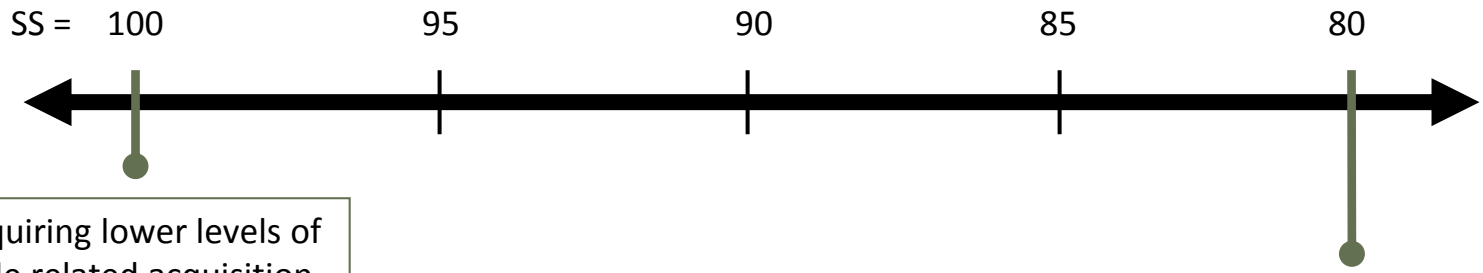
3. *Test performance of ELLs is moderated by the degree to which a given test relies on or requires age- or grade-expected English language development and the acquisition of incidental acculturative knowledge.*

Research Foundations for ELL Evaluation

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



Subtests can be arranged from high to low in accordance with the mean values reported by empirical studies for ELLs



Tests requiring lower levels of age/grade related acquisition of culture and language result in higher mean scores

Tests requiring higher levels of age/grade related acquisition of culture and language result in lower mean scores

Research Foundations for ELL Evaluation

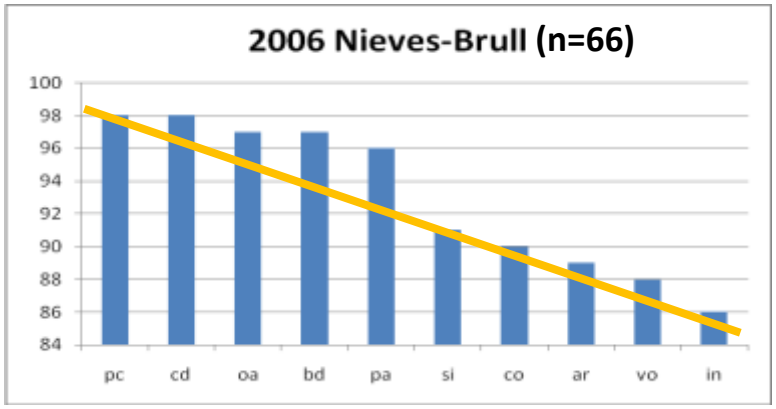
Principle 3: ELL performance is moderated by linguistic/acculturative variables

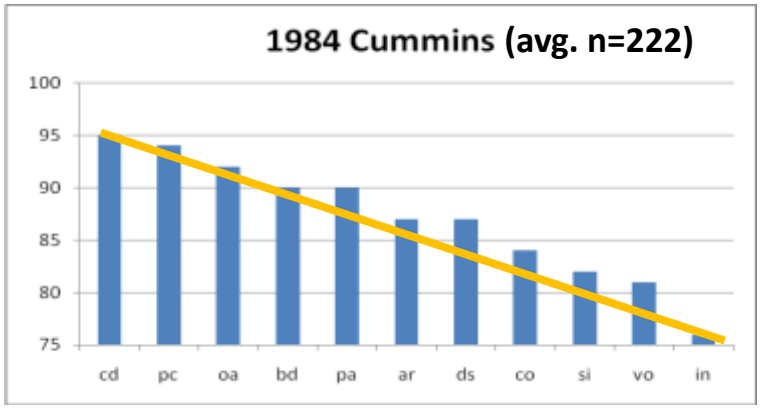
	Hispanic Group (Mercer) (1972)	Hispanic Group (Vukovich & Figueroa) (1982)	ESL Group (Cummins) (1982)	Bilingual Group (Nieves-Brull) (2006)
Subtest Name	**Mean SS**	**Mean SS**	**Mean SS**	**Mean SS**
Information	7.5	7.8	5.1	7.2
Vocabulary	8.0	8.3	6.1	7.5
Similarities	7.6	8.8	6.4	8.2
Comprehension	7.8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

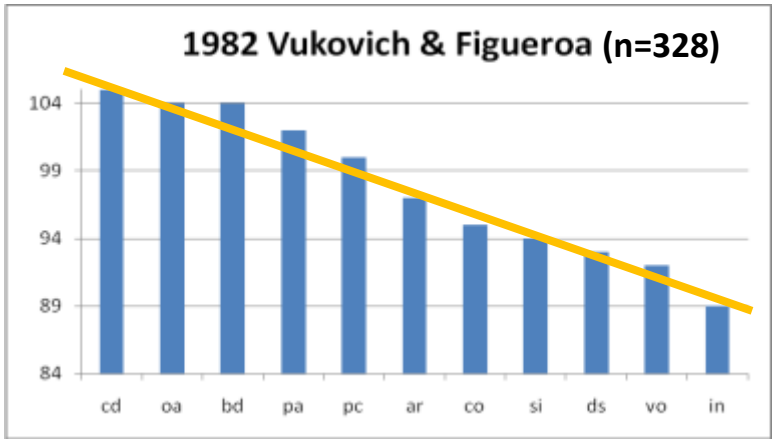
**Data for this subtest were not reported in the study.*

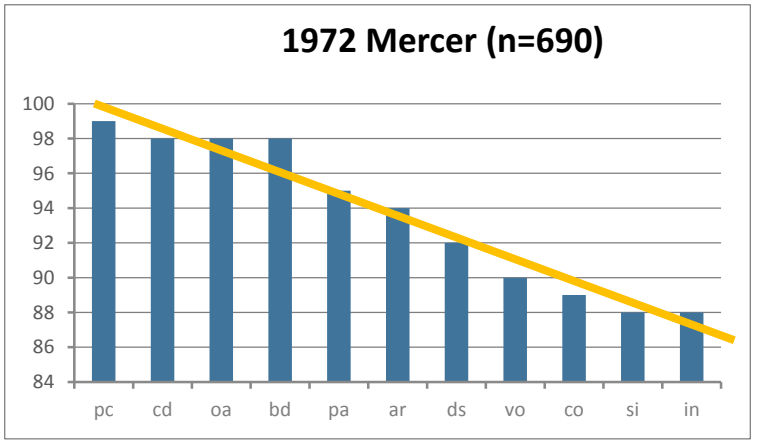
Research Foundations for ELL Evaluation

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



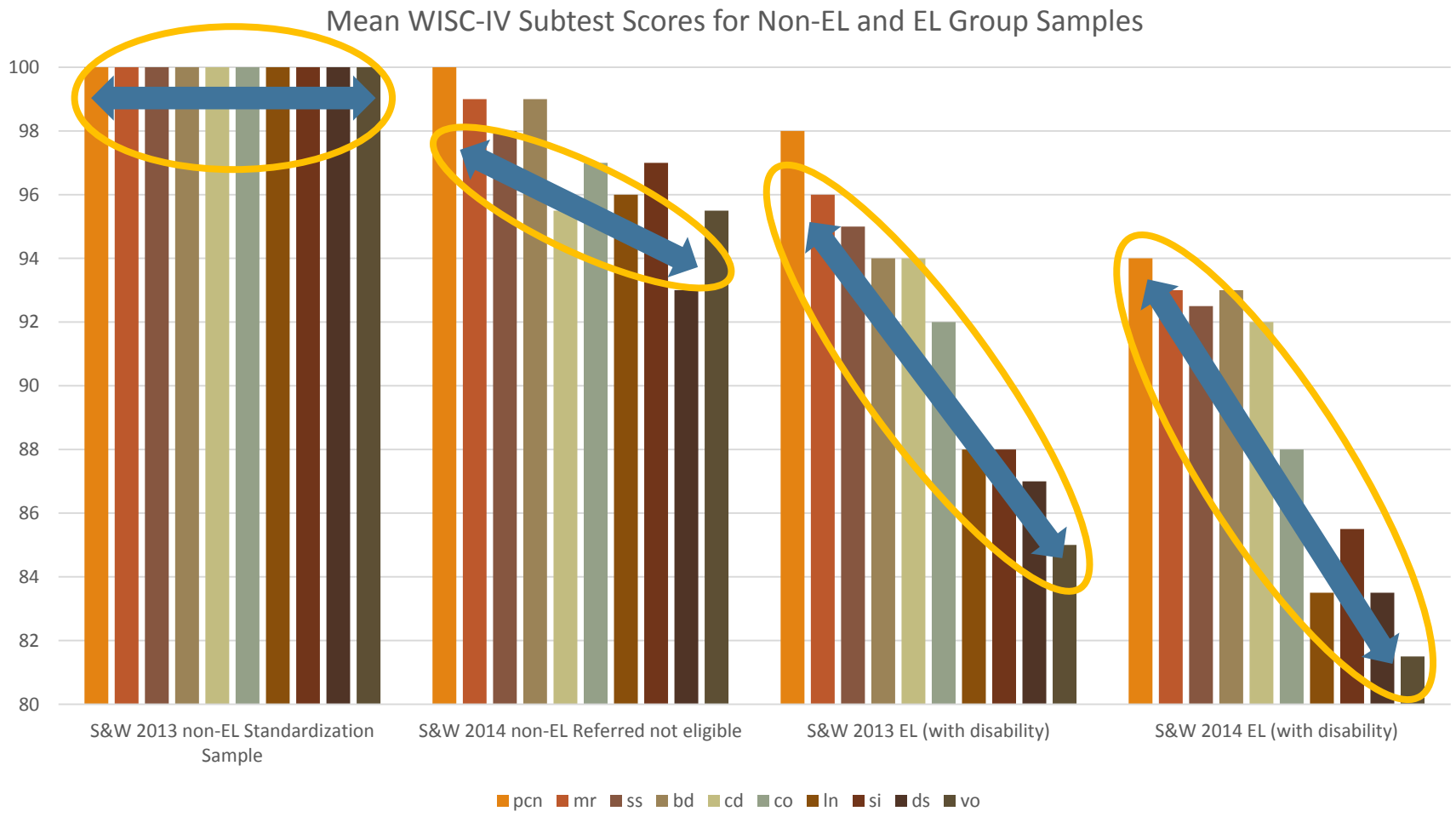






Research Foundations for ELL Evaluation

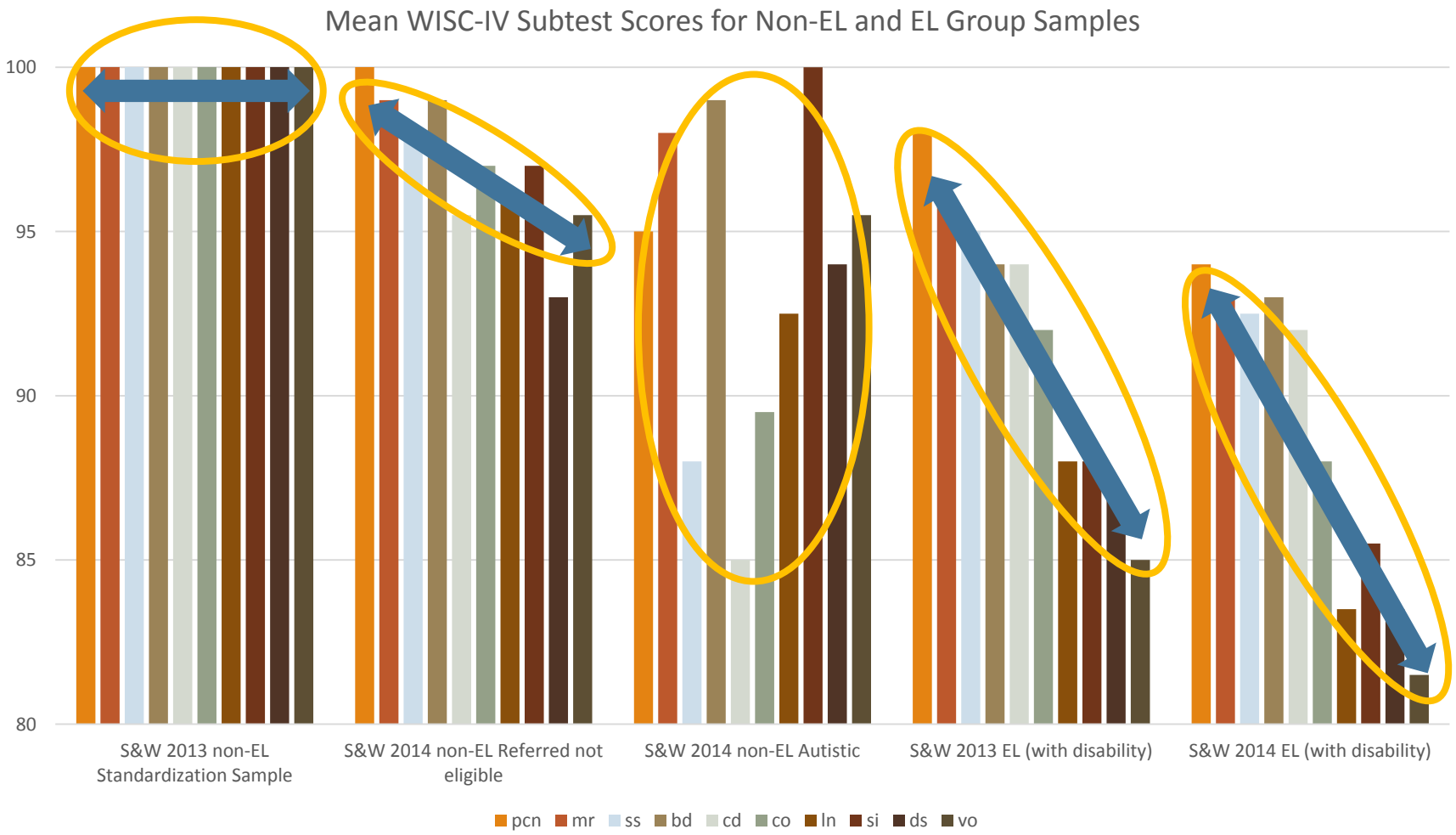
Principle 3: ELL performance is moderated by linguistic/acculturative variables



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

Research Foundations for ELL Evaluation

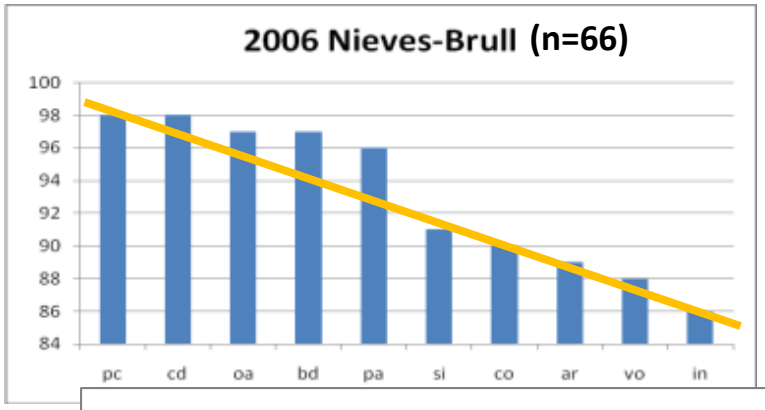
Principle 3: ELL performance is moderated by linguistic/acculturative variables

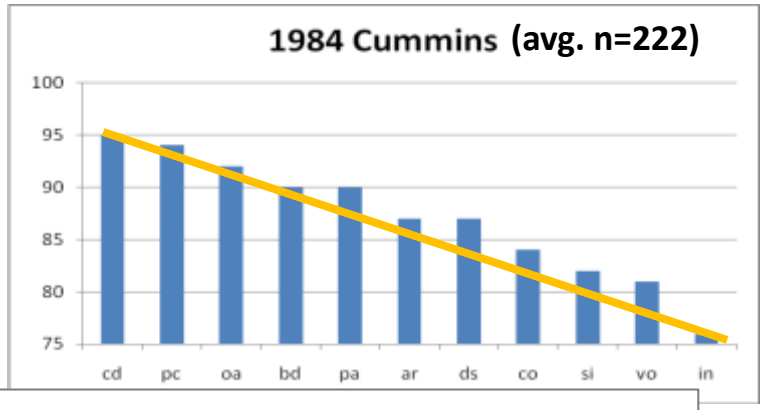


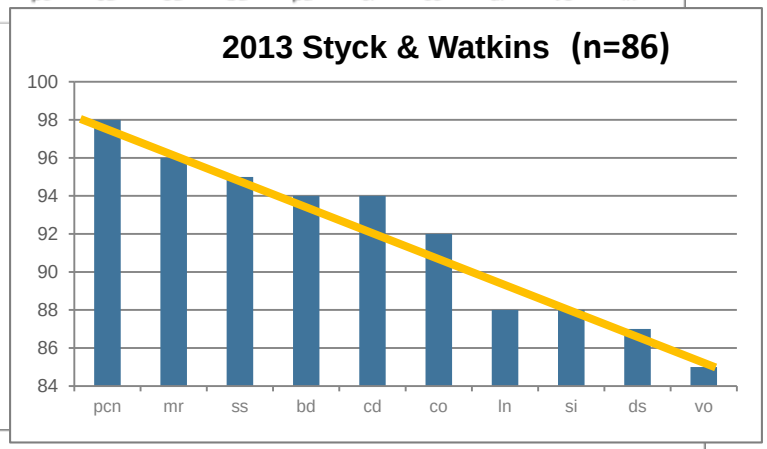
Soures: Styck, K. M. & Watkins, M. W. (2013). Diagnostic Utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition Among Referred Students. School Psychology Review, 42(4), 367-382. and Styck, K. M. & Watkins, M. W. (2014). Discriminant Validity of the WISC-IV Culture-Language Interpretive Matrix. Contemporary School Psychology , 18, 168-188.

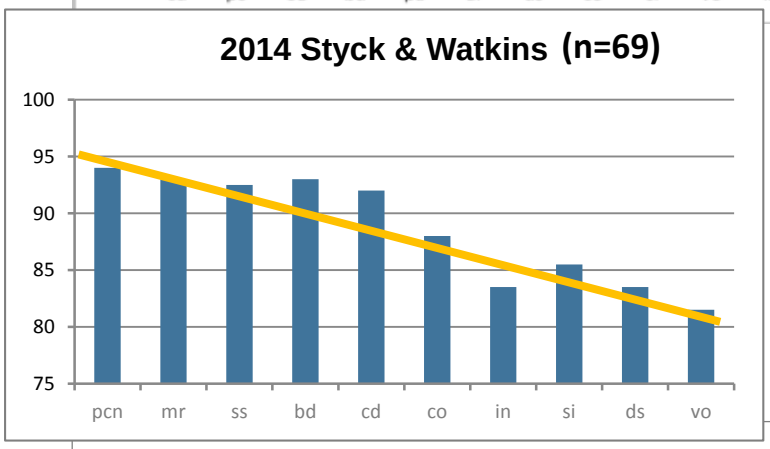
Research Foundations for ELL Evaluation

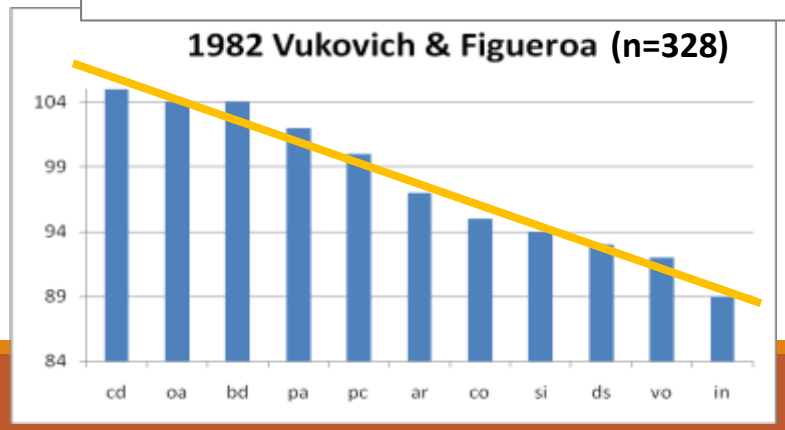
Principle 3: ELL performance is moderated by linguistic/acculturative variables

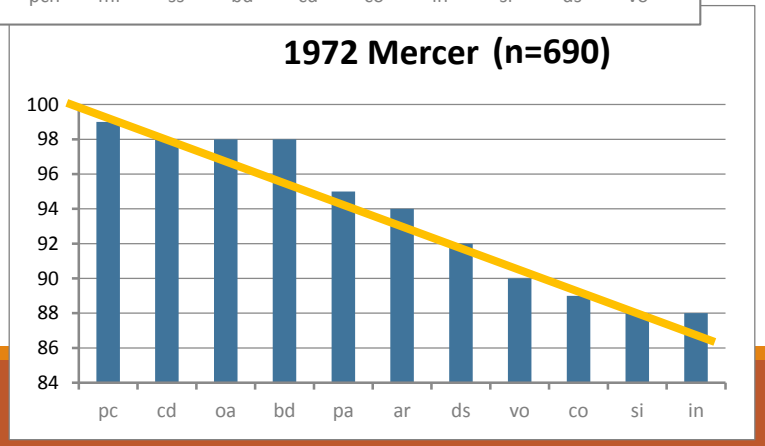








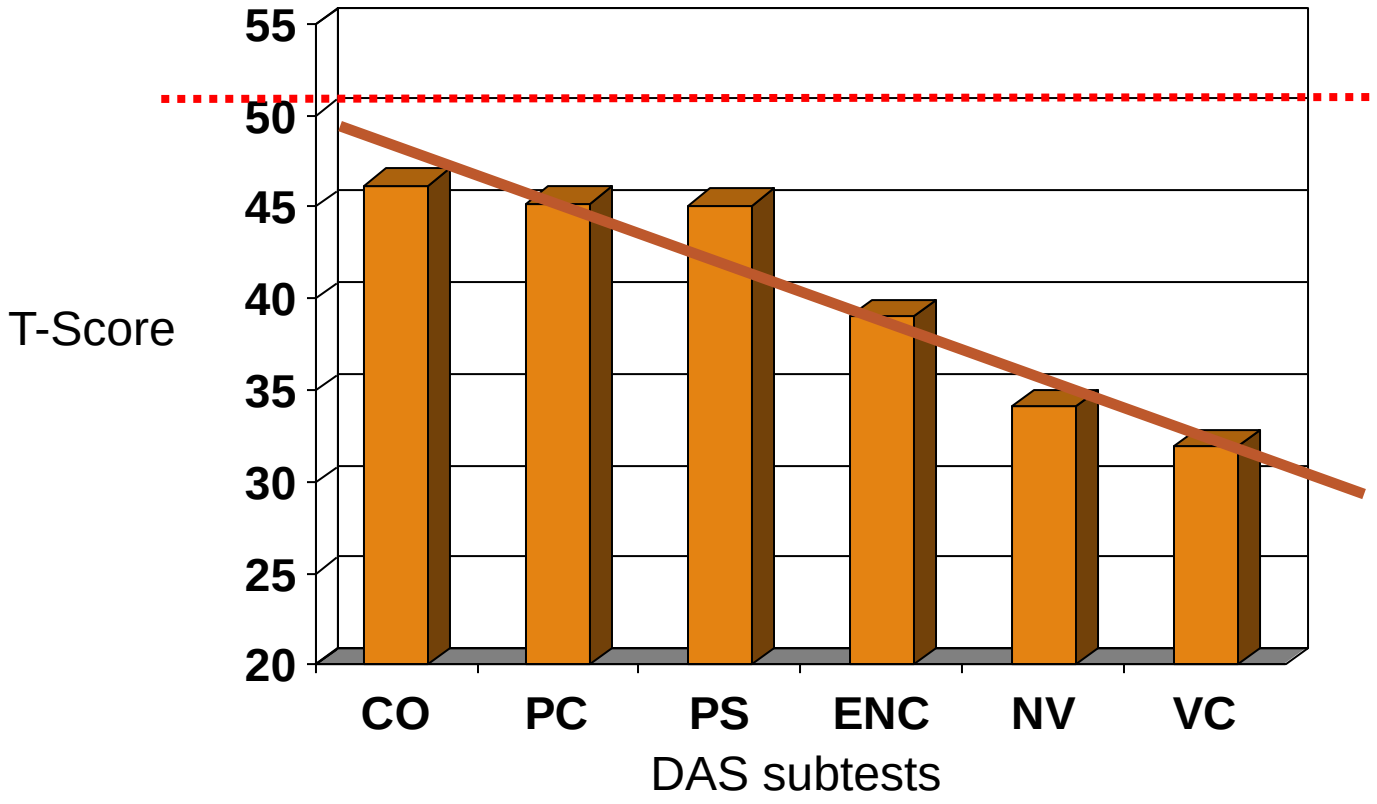




Research Foundations for ELL Evaluation

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

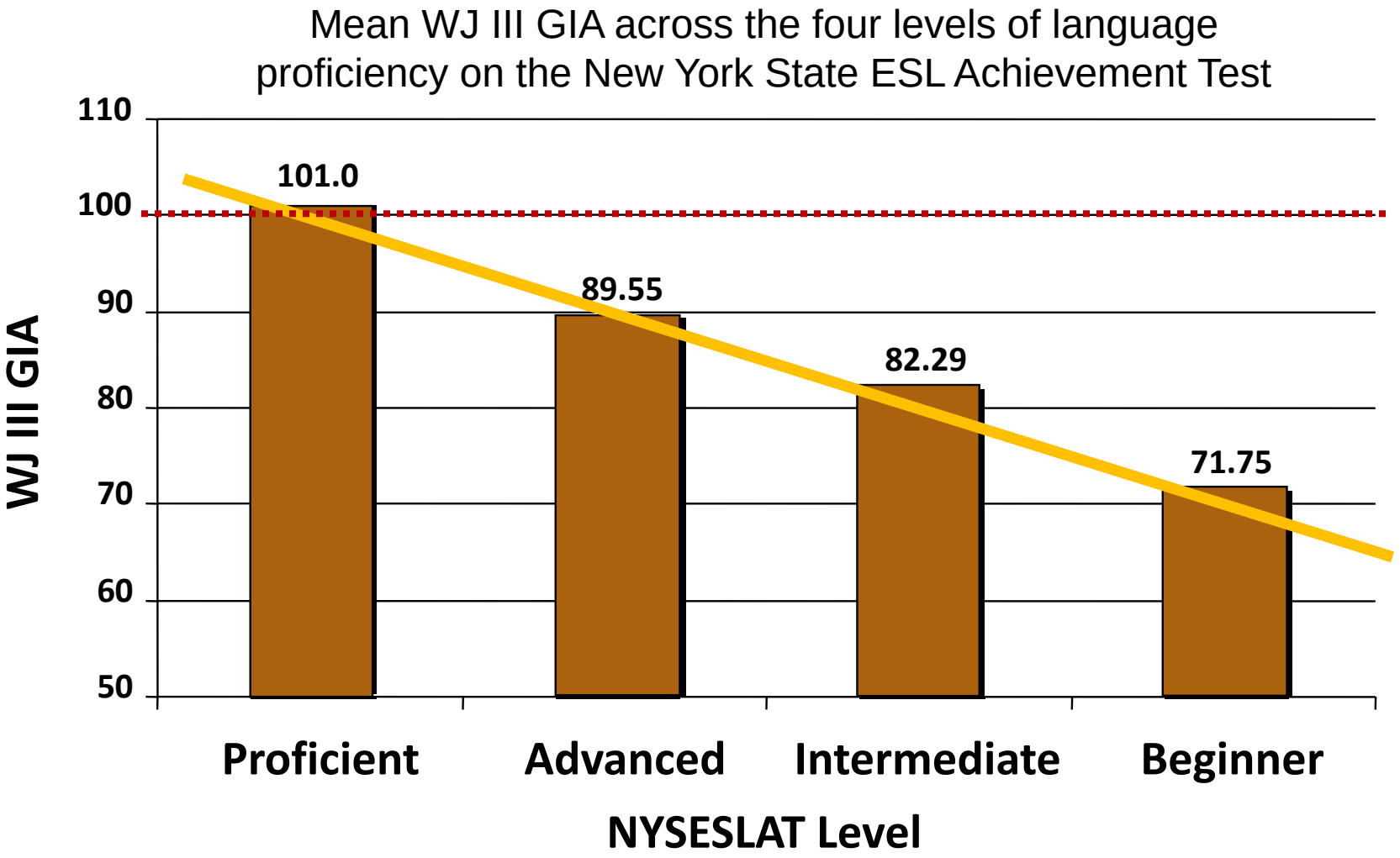
Mean subtest scores across six Differential Ability Scale (DAS) subtests in a pre-school sample of English Language Learners



Source: Aguerra, F., Terjesen, M., Flanagan, D. P., & Ortiz, S. O. (2007). unpublished data.

Research Foundations for ELL Evaluation

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



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

Research Foundations for ELL Evaluation

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

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

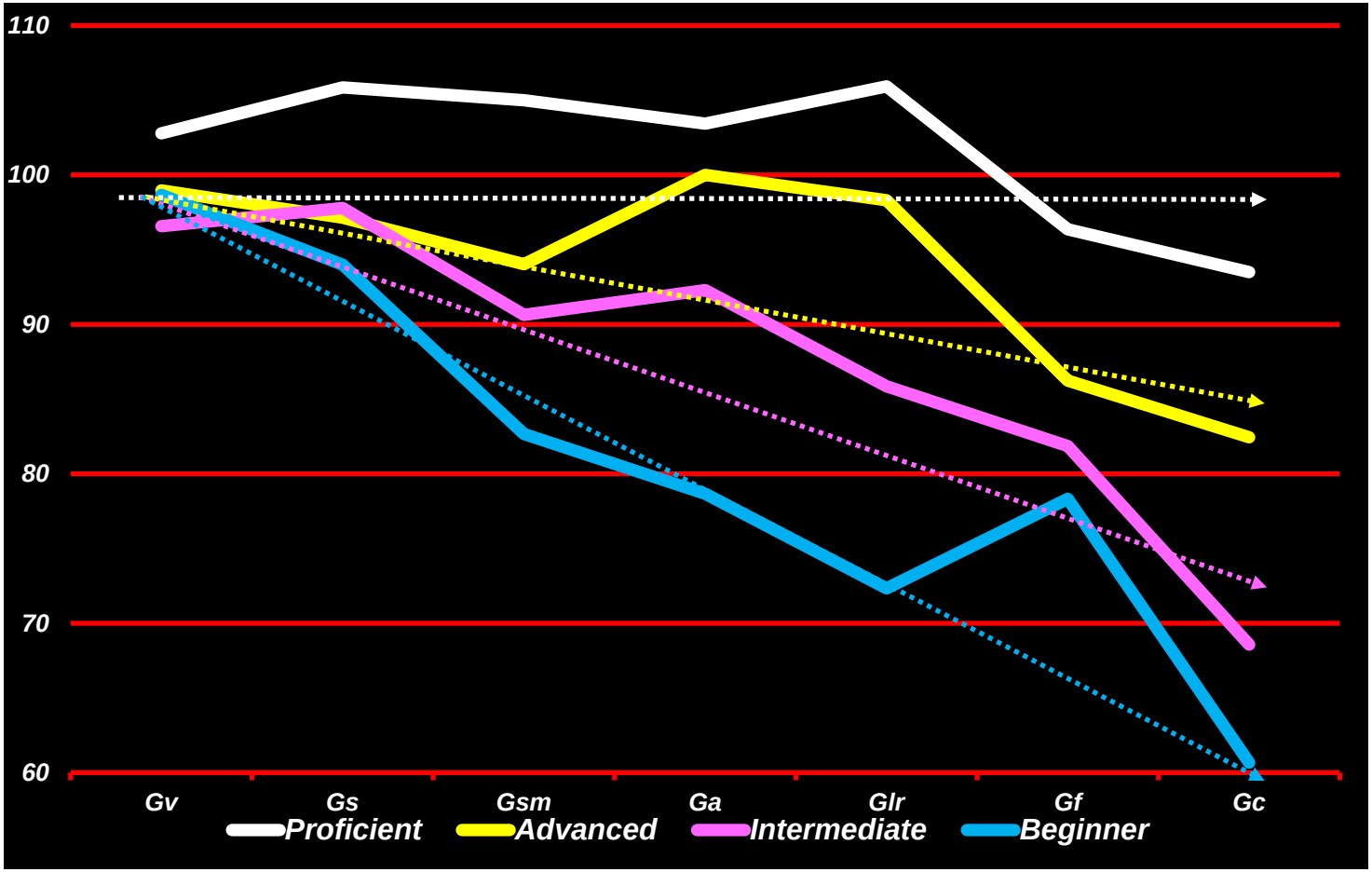
	Variance explained		
Individual test	7-10	11-14	15-18
Verbal Comprehension	.79[c]	.86[c]	.81[c]
General Information	.71[c]	.85[c]	.86[c]
Concept Formation	.67[c]	.71[c]	.67[c]
Visual–Auditory Learning	.40[b]	.37[b]	.41[b]
Delayed Recall Visual–Auditory Learning	.39[b]	.32[b]	.37[b]
Analysis Synthesis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

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

Research Foundations for ELL Evaluation

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

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

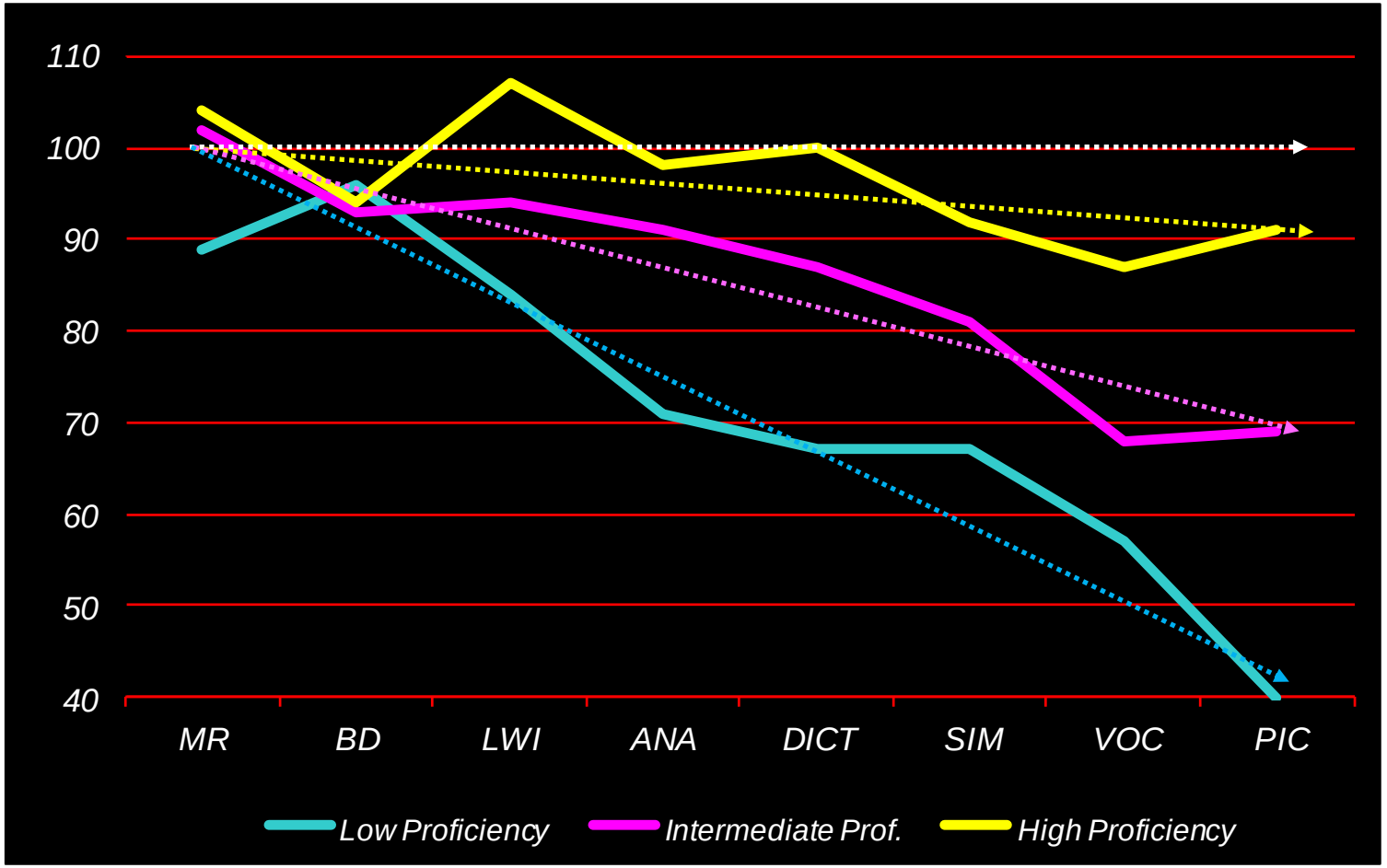


Source: Sotelo-Dynega, M., Ortiz, S.O., Flanagan, D.P., Chaplin, W. (2013). English Language Proficiency and Test Performance: Evaluation of bilinguals with the Woodcock-Johnson III Tests of Cognitive Ability. Psychology in the Schools, Vol 50(8), pp. 781-797.

Research Foundations for ELL Evaluation

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

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



Source: Dynda, A.M., Flanagan, D.P., Chaplin, W., & Pope, A. (2008), unpublished data..

Main Threats to Test Score Validity for ELLs

NO BIAS

- ***Test items***
  (content, novelty)
- ***Test structure***
  (sequence, order, difficulty)
- ***Test reliability***
  (measurement error/accuracy)
- ***Factor structure***
  (theoretical structure, relationship of variables to each other)
- ***Predictive Validity***
  (correlation with academic success or achievement)

BIAS

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

When a test measures an unintended variable…

- ***Incorrect Interpretation***
  (undermines accuracy of evaluative judgments and meaning assigned to scores)

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

Main Threats to Test Score Validity for ELLs

Acculturative Knowledge Acquisition – Not Race or Ethnicity

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

Salvia & Ysseldyke, 1991

Developmental Language Proficiency – Not Language Dominance

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

Lohman, Korb & Lakin, 2008

Processes and Procedures for Addressing Test Score Validity

IX. REDUCE BIAS IN TRADITIONAL TESTING PRACTICES

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

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

Processes and Procedures for Addressing Test Score Validity

ISSUES IN MODIFIED METHODS OF EVALUATION

Modified and Altered Assessment:

- *often referred to as "testing the limits" where the alteration or modification of test items or content, mediating task concepts prior to administration, repeating instructions, accepting responses in either language, and eliminating or modifying time constraints, etc., are employed in efforts to help the examinee perform to the best of their ability*
- *any alteration of the testing process violates standardization and effectively invalidates the scores and precludes interpretation or assignment of meaning*
- *use of a translator/interpreter for administration helps overcome the language barrier but is also a violation of standardization and undermines score validity, even when the interpreter is highly trained and experienced; tests are not usually normed in this manner*
- *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***
- *alterations or modifications are perhaps most useful in deriving qualitative information—observing behavior, evaluating learning propensity, evaluating developmental capabilities, analyzing errors, etc.*
- *a recommended procedure would be to administer tests in a standardized manner first, which will potentially allow for later interpretation, and then consider any modifications or alterations that will further inform the referral questions*

Processes and Procedures for Addressing Test Score Validity

ISSUES IN NONVERBAL METHODS OF EVALUATION

Language Reduced Assessment:

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

Processes and Procedures for Addressing Test Score Validity

ISSUES IN DOMINANT LANGUAGE EVALUATION: Native language

Native Language Assessment (L1):

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

ELL Test Performance: Esparza Brown Study

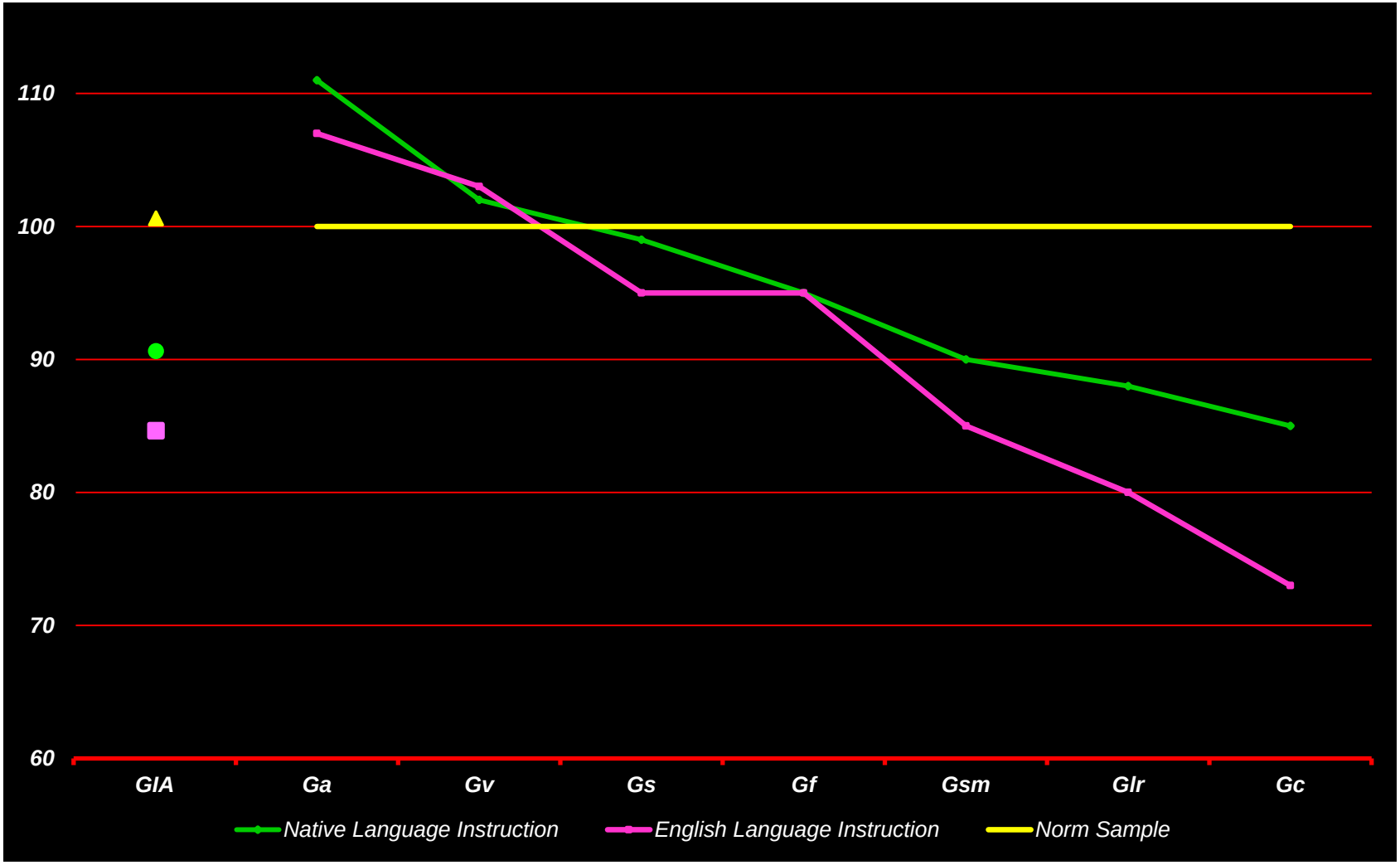
Comparison of Order of Means for WJ III and Bateria III Classifications*

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

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

ELL Test Performance: Esparza Brown Study

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



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

Processes and Procedures for Addressing Test Score Validity

ISSUES IN DOMINANT LANGUAGE EVALUATION: English

English Language Assessment (L2):

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

Processes and Procedures for Addressing Test Score Validity

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

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

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

Comparison of Methods for Addressing Main Threats to Validity

Evaluation Method	Norm sample representative of bilingual development	Measures full range of ability constructs	Does not require bilingual evaluator	Adheres to the test's standardized protocol	Substantial research base on bilingual performance
Modified or Altered Assessment	✗	✓	✓	✗	✗
Reduced-language Assessment	✗	✗	✓	✓	✗
Dominant Language Assessment – L1 (native language)	✗	✓	✗	✓	✗
Dominant Language Assessment – L2 (English)	✗	✓	✓	✓	✓

Addressing issues of fairness with respect to norm sample representation is an issue of validity and dependent on a sufficient research base.

Evaluating and Defending Construct ELL Test Score Validity

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

"DIFFERENCE vs. DISORDER?"

Simply absolving oneself from responsibility of doing so via wording such as, "all scores should be interpreted with extreme caution" does not in any way provide a defensible argument regarding the validity of obtained test results and does not permit interpretation.

At present, the only manner in which test score validity can be evaluated or established is via use of the existing research on the test performance of ELLs as reflected in the degree of "difference" the student displays relative to the norm samples of the tests being used, particularly for tests in English. This is the sole purpose of the C-LIM.

Foundational Research Principles of the Culture-Language Interpretive Matrix

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

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

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

Because the basic research principles underlying the C-LIM are well supported, their operationalization within the C-LIM provides a substantive evidentiary base for evaluating the test performance of English language learners.

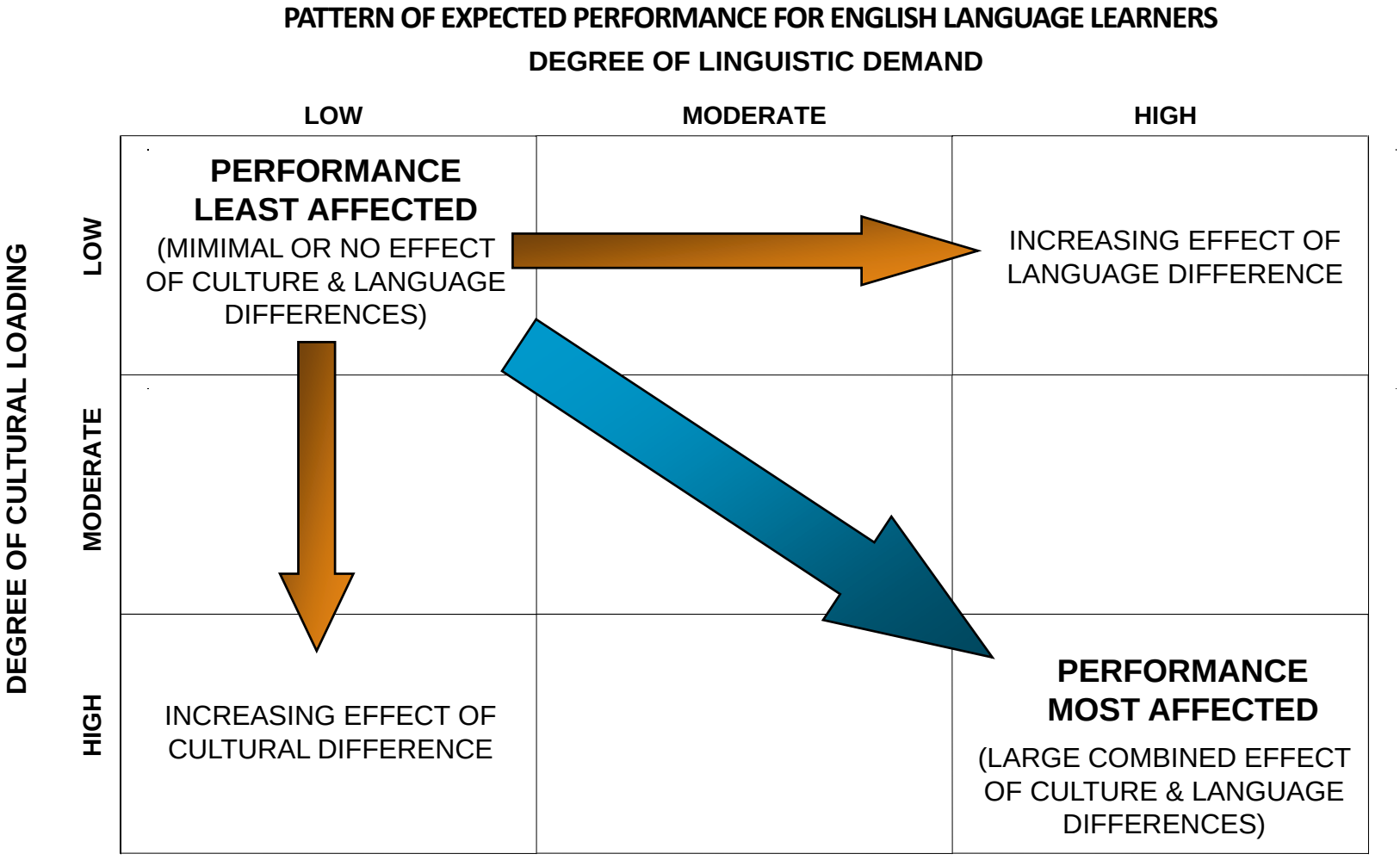
- This does not mean, however, that it cannot be improved. Productive research on EL test performance can assist in making any necessary "adjustments" to the order of the means as arranged in the C-LIM.
- Likewise, as new tests come out, new research is needed to determine the relative level of EL performance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

The Culture-Language Interpretive Matrix (C-LIM)
Addressing test score validity for ELLs

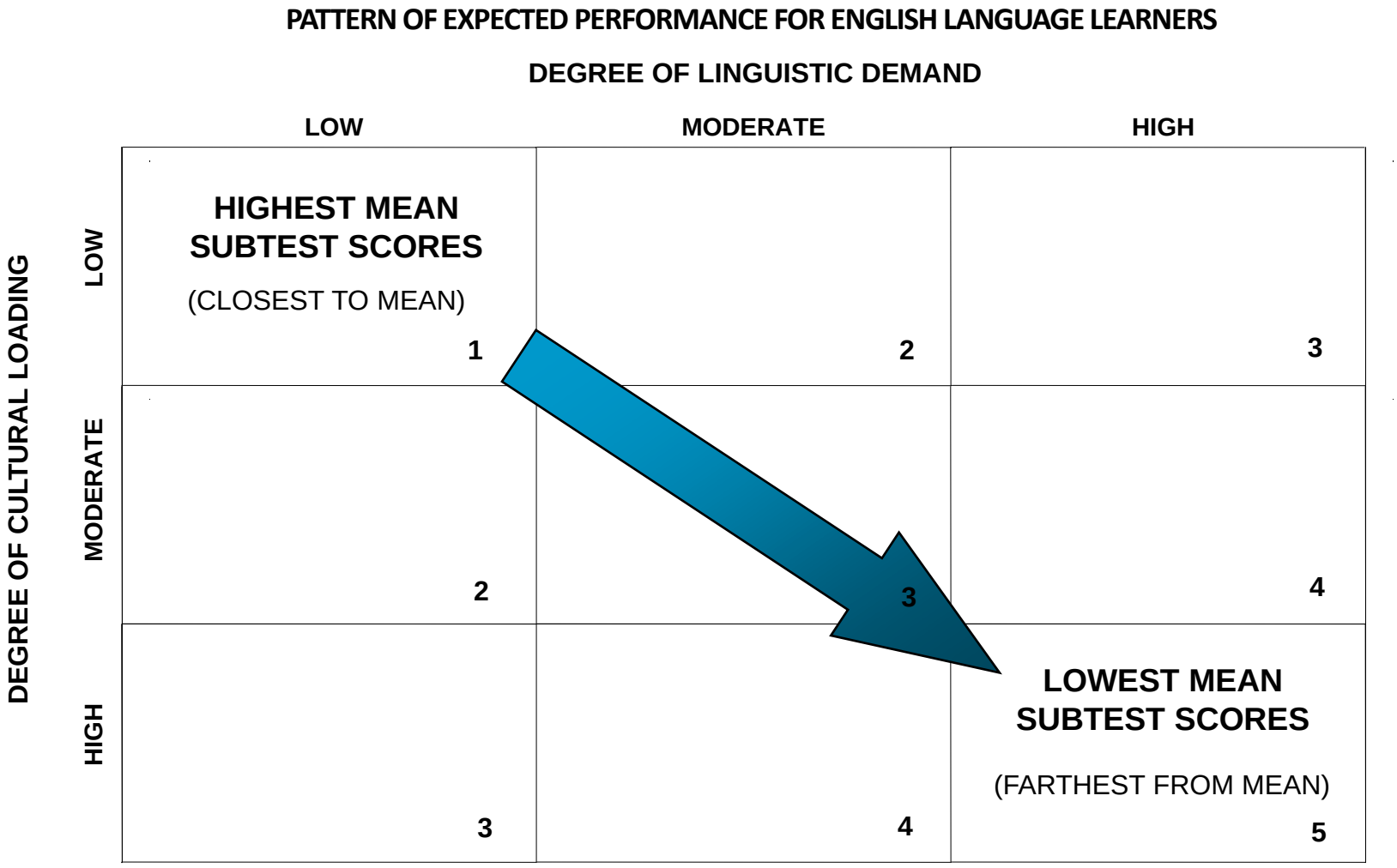
Translation of Research into Practice

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

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



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



The Culture-Language Interpretive Matrix (C-LIM)

Important Facts for Use and Practice

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

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

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

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

Practical Considerations for Addressing Validity in Evaluation Procedures for SLD with ELLs

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

Practical Considerations for Addressing Validity in Evaluation Procedures for SLD with ELLs

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

- *Test in English first and if all test scores indicate strengths (average or higher) a disability is not likely and thus no further testing is necessary*

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

This approach provides the most efficient process and best use of available resources for evaluation since it permits ANY evaluator to begin and sometimes complete the testing without being bilingual or requiring assistance.

In addition, this approach is IDEA compliant and consistent with the specification that assessments "be provided and administered in the language and form most likely to yield accurate information" because it relies on an established body of research to guide examination of test score validity and ensures that that the results upon which decisions are based are in fact accurate.

A Recommended Best Practice Approach for Using Tests with ELLs

Step 1. Assessment of Bilinguals – validate all areas of performance (exclusion of cultural/linguistic factors)

- *Select or create an appropriate battery that is comprehensive and responds to the needs of the referral concerns, irrespective of language differences*
- *Administer all tests in standardized manner first in English only with no modifications*
- *Score tests and plot them for analysis via the C-LIM*
- *If analysis indicates expected range and pattern of decline, scores are invalid due to cultural and linguistic factors that cannot be excluded as primary reason for poor academic performance*
- *If analysis does not indicate expected range or pattern of decline, apply XBA (or other) interpretive methods to determine specific areas of weakness and difficulty and continue to Step 2*

Step 2. Bilingual Assessment – validate suspected areas of weakness (cross-language confirmation of deficit areas)

- *Review results and identify areas of suspected weakness or difficulty:*
    a. *For Gc only, evaluate weakness according to high/high cell in C-LIM or in context of other data and information*
    b. *For all other abilities, evaluate weakness using standard classifications (e.g., SS < 90)*
- *Except for Gc, re-test all other areas of suspected weakness using native language tests*
- *For Gc only:*
    a. *If the high/high cell in C-LIM is within/above expected range, consider Gc a strength and assume it is at least average, thus re-testing is not necessary*
    b. *If the high/high cell in C-LIM is below expected range, re-testing of Gc in the native language is recommended*
- *Administer native language tests or conduct re-testing using one of the following methods:*
    a. *Native language test administered in the native language (e.g., WJ III/Bateria III or WISC-IV/WISC-IV Spanish)*
    b. *Native language test administered via assistance of a trained interpreter*
    c. *English language test translated and administered via assistance of a trained interpreter*
- *Administer tests in manner necessary to ensure full comprehension including use of any modifications and alterations necessary to reduce barriers to performance, while documenting approach to tasks, errors in responding, and behavior during testing, and analyze scores both quantitatively and qualitatively to confirm and validate areas as true weaknesses*
- *Except for Gc, if a score obtained in the native language validates/confirms a weakness score obtained in English (both SS < 90), use/interpret the score obtained in English as a weakness*
- *If a score obtained in the native language invalidates/disconfirms a weakness score obtained in English (native SS $\geq$ 90), consider it as a strength and assume that it is at least in the average range*
- *Scores for Gc obtained in the native language and in English can only be interpreted relative to developmental and educational experiences of the examinee in each language and only as compared to others with similar developmental experiences*

The Culture-Language Test Classifications and Interpretive Matrix: Summary and Conclusions

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

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

…the C-LTC and C-LIM can be of practical value in helping establish credible and defensible validity for test data, thereby decreasing the potential for biased and discriminatory interpretation. Taken together with other assessment data, the C-LTC and C-LIM assist practitioners in answering the most basic question in ELL assessment:

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

Nondiscriminatory Assessment and Standardized Testing

"Probably no test can be created that will entirely eliminate the influence of learning and cultural experiences. The test content and materials, the language in which the questions are phrased, the test directions, the categories for classifying the responses, the scoring criteria, and the validity criteria are all culture bound."

- *Jerome M. Sattler, 1992*

Assessment of English Language Learners - Resources

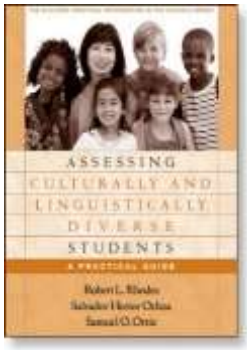
BOOKS:

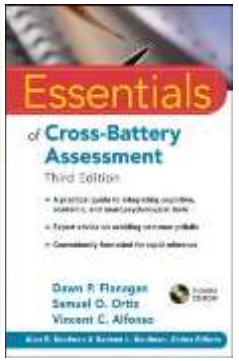
Rhodes, R., Ochoa, S. H. & Ortiz, S. O. (2005). Comprehensive Assessment of Culturally and Linguistically Diverse Students: A practical approach. New York: Guilford.

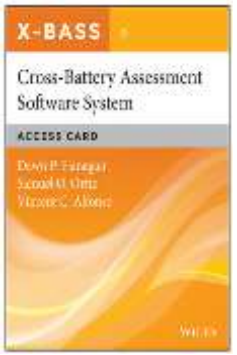
Flanagan, D. P., Ortiz, S.O. & Alfonso, V.C. (2013). Essentials of Cross-Battery Assessment, Third Edition. New York: Wiley & Sons, Inc.

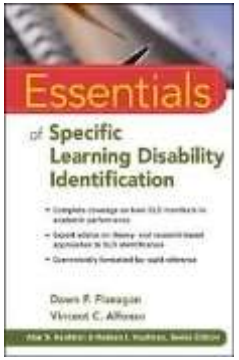
Flanagan, D.P. & Ortiz, S.O. (2012). Essentials of Specific Learning Disability Identification. New York: Wiley & Sons, Inc.

Ortiz, S. O., Flanagan, D. P. & Alfonso, V. C. (2015). Cross-Battery Assessment Software System (X-BASS v1.0). New York: Wiley & Sons, Inc.









ONLINE:

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

Cross-Battery XBA Assessment