English Learners and Assessment:

A developmental framework for promoting equitable evaluation.

Confederation of Oregon School Administrators
40th Annual Conference
Seaside, OR

June 18, 2014

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

Cultural and Linguistic Issues in Early Testing

The newly transformed Binet Scales were thought to provide a psychometric tool that could precisely measure intelligence independent of other factors. To maintain this perspective required unquestioned belief that:

- *Intelligence was genetic, innate, static, immutable, and largely unalterable by experience, opportunity, or environment*
- *Whether or not you fully comprehended or spoke English did not significantly affect testing*
- *Familiarity with and knowledge of U.S. culture had no bearing on intelligence test performance*
- *Being raised in another culture or having different cultural experiences was irrelevant*

"Intelligence is what intelligence tests measures" (Boring, 1923), and that means that "you are what the test says you are."

- *Being bilingual was itself the problem because it resulted in a "mental handicap" measured accurately by poor performance on intelligence tests and thus substantiating its detrimental influence*

Cultural and Linguistic Issues in Early Testing

Much of the these original perspectives and ideas regarding the meaning of test results, particularly with respect to cultural and linguistic differences, remain embedded in various ways in present day tests:

Very Superior	evolved from	Precocious
Superior		Superior
High Average		Normal
Average		Borderline
Low Average		Moron
Borderline		Imbecile
Deficient		Idiot

In 1974, the following question was asked on the WISC-R:

- *Who discovered America?*

In 1991, with "attention" to issues regarding cultural fairness, the same question on the WISC-III was "changed" to:

- *Who was Christopher Columbus?*

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

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

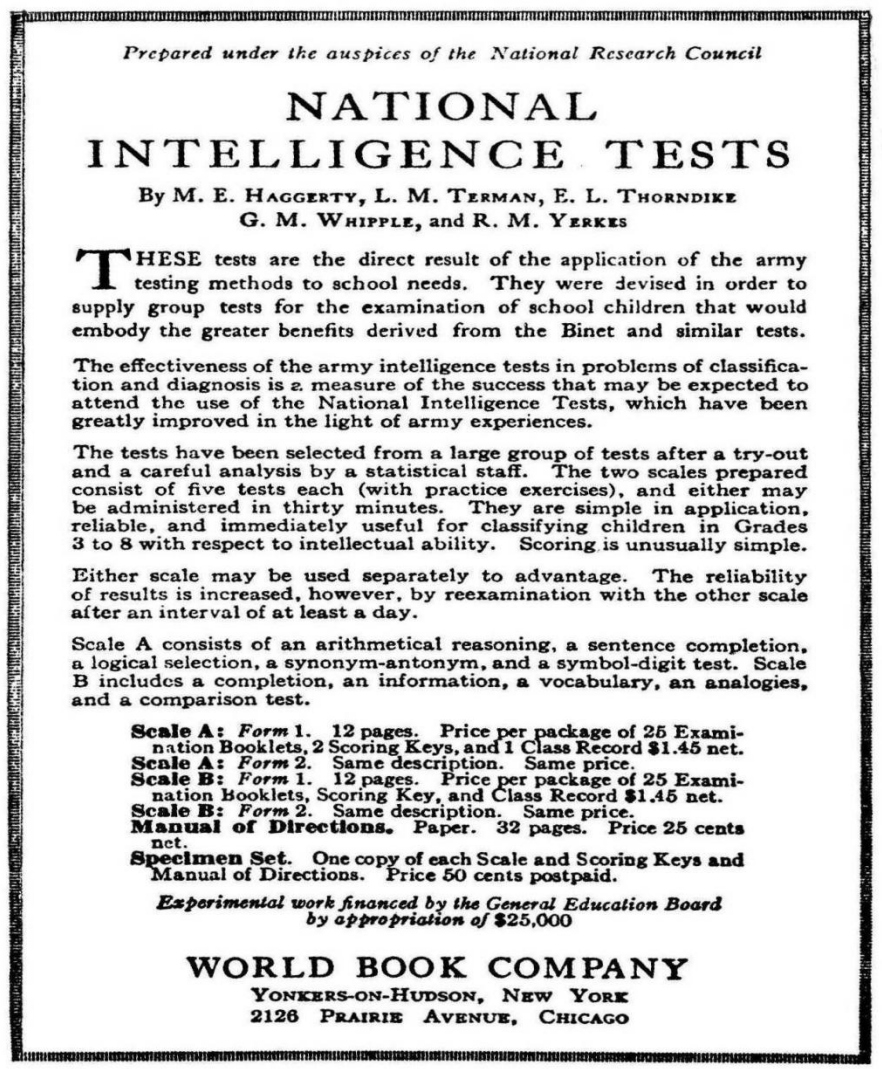
- *The testing of newly arrived immigrants at Ellis Island*

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.

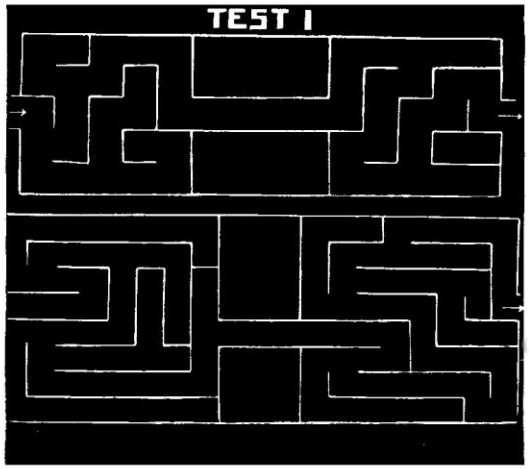
Either scale may be used separately to advantage. The reliability of results is increased, however, by reexamination with the other scale after an interval of at least a day.

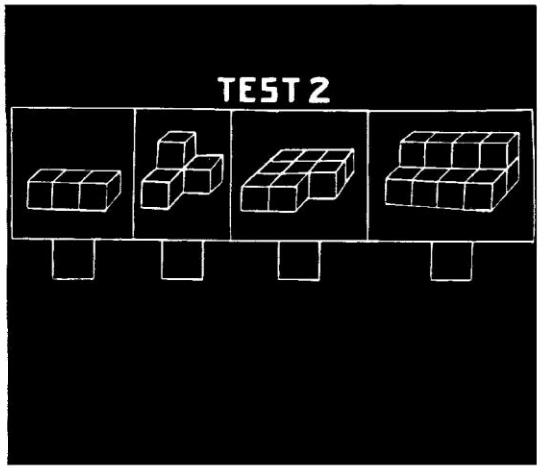
Scale A consists of an arithmetical reasoning, a sentence completion, a logical selection, a synonym-antonym, and a symbol-digit test. Scale B includes a completion, an information, a vocabulary, an analogies, and a comparison test.

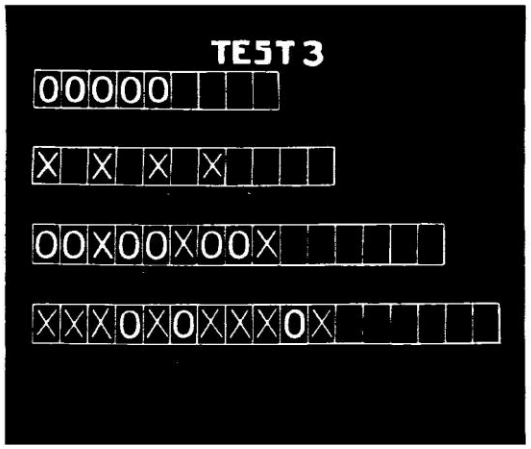
Scale A: *Form* 1. 12 pages. Price per package of 25 Examination Booklets, 2 Scoring Keys, and 1 Class Record $1.45 net.
Scale A: *Form* 2. Same description. Same price.
Scale B: *Form* 1. 12 pages. Price per package of 25 Examination Booklets, Scoring Key, and Class Record $1.45 net.
Scale B: *Form* 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.

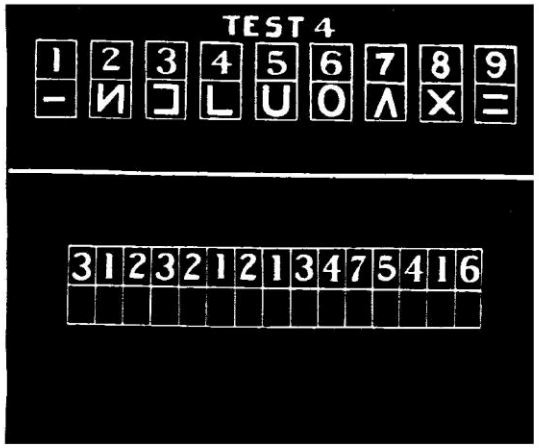
Experimental work financed by the General Education Board by appropriation of $25,000

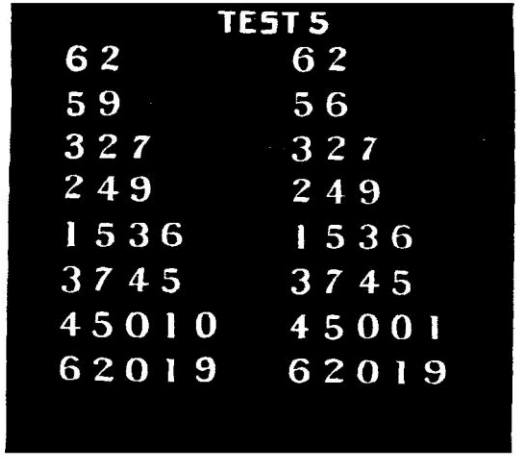
WORLD BOOK COMPANY
Yonkers-on-Hudson, New York
2126 Prairie Avenue, Chicago

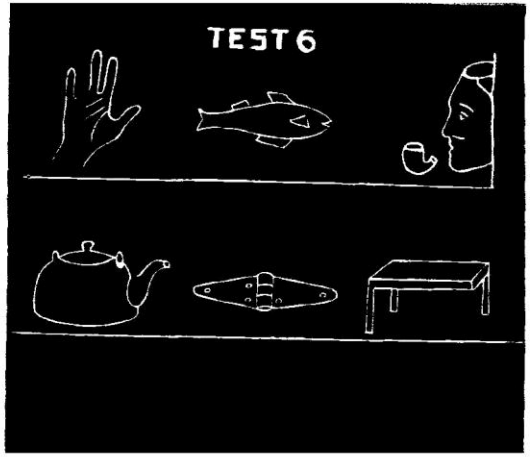


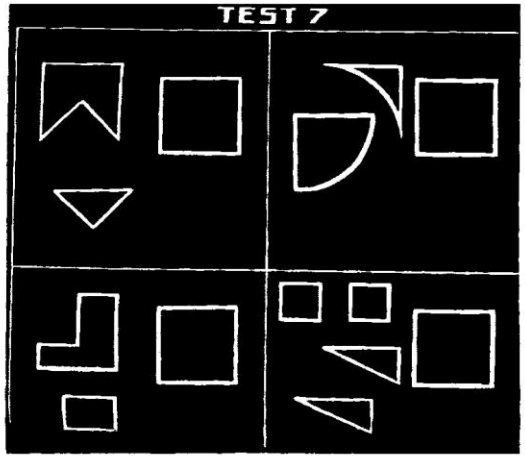




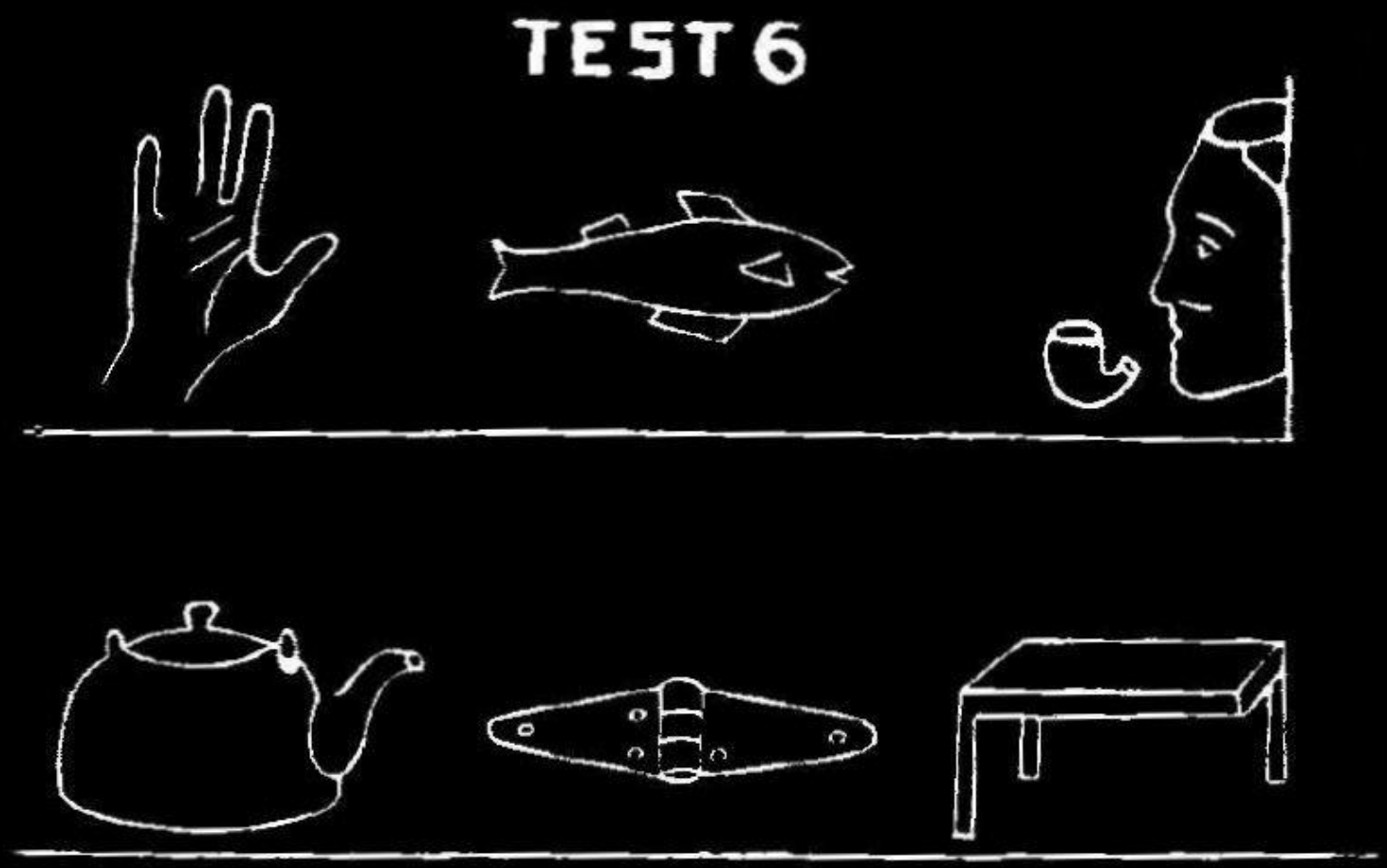






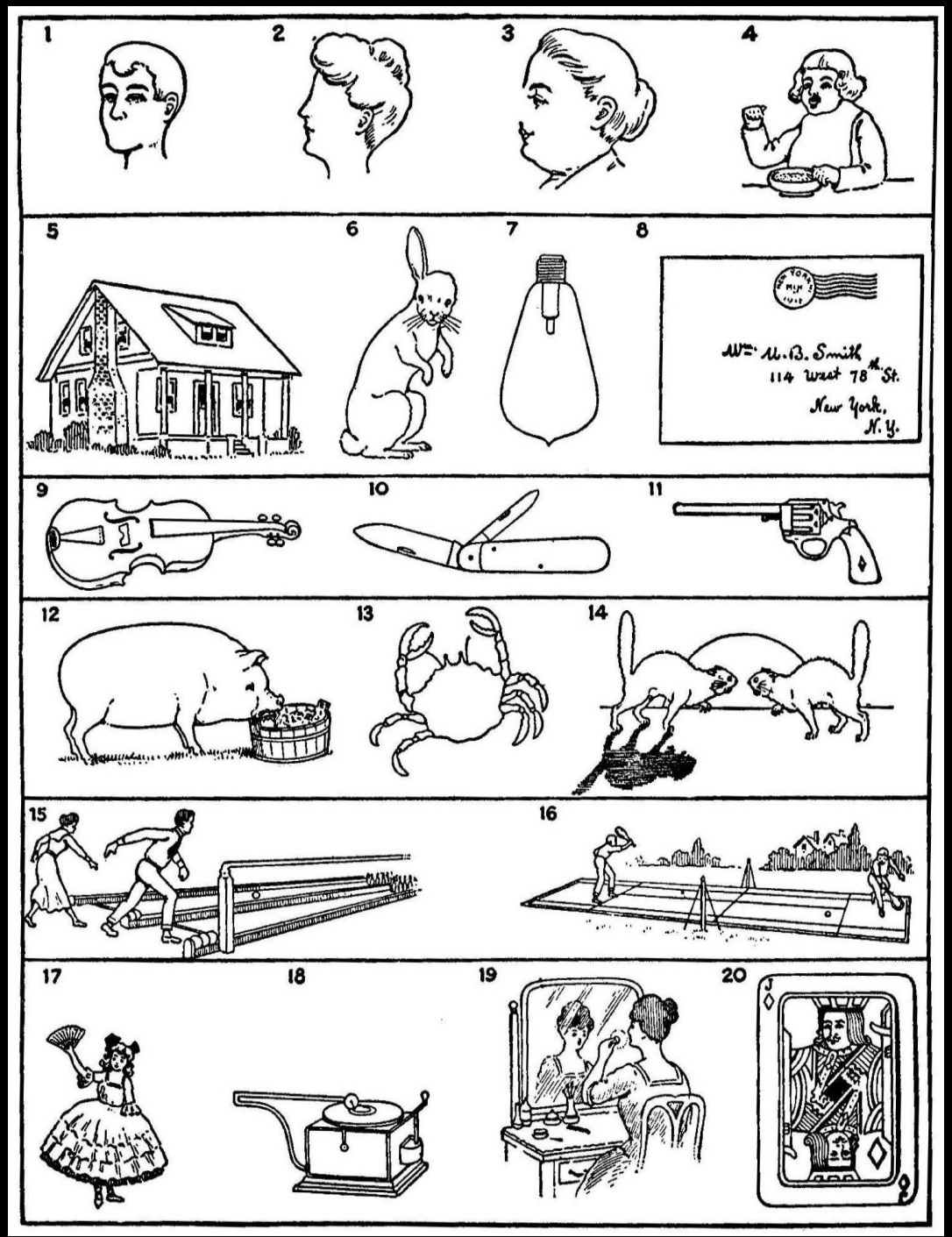


The blackboard demonstrations for seven parts of the Beta Test. From Yerkes, 1921.



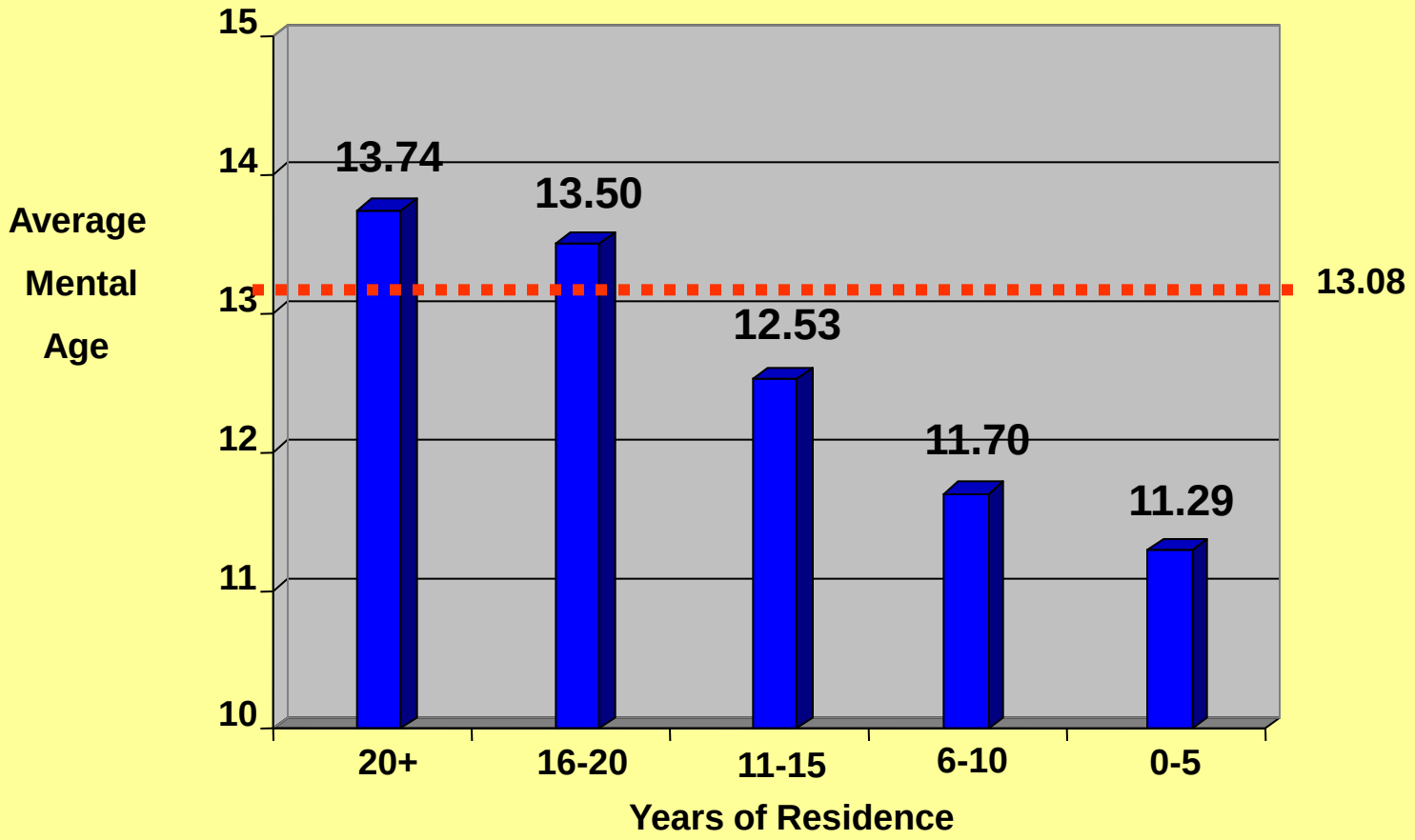
Instructional Items from Test 6 of the Army Beta Test.

Part six of examination Beta for testing innate intelligence.



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

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

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

Bilingualism and Testing

- *Interpretation: New immigrants are inferior*

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

Brigham, 1923

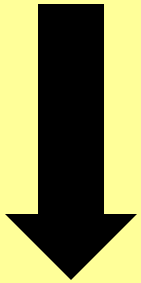
The Nature of Bias in Tests and Testing: It's not what you think.

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*

The Nature of Bias in Tests and Testing

"The fact that the Mexican group is very similar to the white in rank order of p values and p decrements on both the PPVT and the Raven, yet has lower scores on the PPVT than on the Raven, suggests that some factor is operating to depress the PPVT performance more or less uniformly for all items and that this factor does not depress Raven performance, at least to the same degree. It seems plausible to suggest that this factor is verbal and may be association with bilingualism in the Mexican group"

Jensen, 1974

"Thus, there is some evidence that a vocabulary test in English may be a biased test of intelligence for Mexican-American's" *Jensen, 1976*

The Nature of Bias in Tests and Testing

- *As long as tests are arranged according to developmental level with increasing difficulty they retain their psychometric properties with respect to measurement accuracy (reliability) and predictive validity.*

- *When experiential circumstances create developmental differences that no longer match up with age expectancies related to the acquisition of language and acculturative knowledge, test performance will be attenuated.*

- *For English learners, the experiential circumstances that lead to age-related disruption in expected developmental typically occur when the learning of English (as a second language) begins at some point other than birth, when age-appropriate education occurs in a language other than the native one, or when formal instruction in any language begins at a point later than it should.*

So What Factors Most Threaten the Validity of Test Performance?

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, p. 276-278.

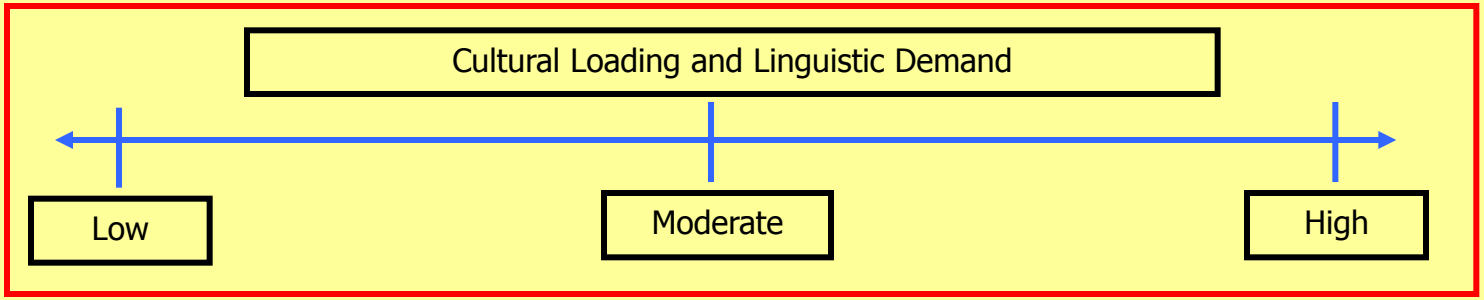
Acquisition of Language and Cultural Knowledge are Developmental Processes Embedded in Tests

- **Tests require age/grade related acquisition of culture (knowledge):**
  - the majority of tests used by psychologists were developed and normed in U.S. and inherently reflect native anthropological content as well as the culturally bound conceptualizations of the test developers themselves. Many tests require specific prior knowledge of, experience with, and even fluency regarding mainstream U.S. culture

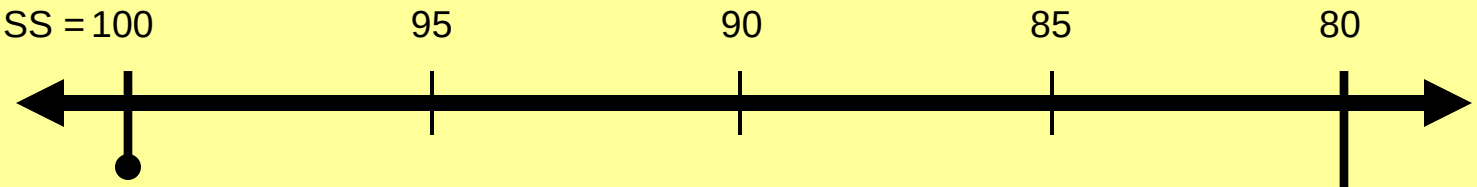
- **Tests require age/grade related acquisition of language (communication):**
  - linguistic factors affect administration, comprehension, responses, and performance on virtually all tests. Even nonverbal tests that reduce oral language requirements continue to rely on effective communication between examiner and examinee in order to measure optimal performance

- **Tests vary on both dimensions:**
  - Tests vary significantly with respect to the degree that they are culturally loaded as well as the degree of language required

Test Performance is Mediated Proportionally by Differences in Developmental Experiences



But test characteristics alone are insufficient to reflect differences rooted in development. Mean values are needed.



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

Acculturative Knowledge and Language Proficiency

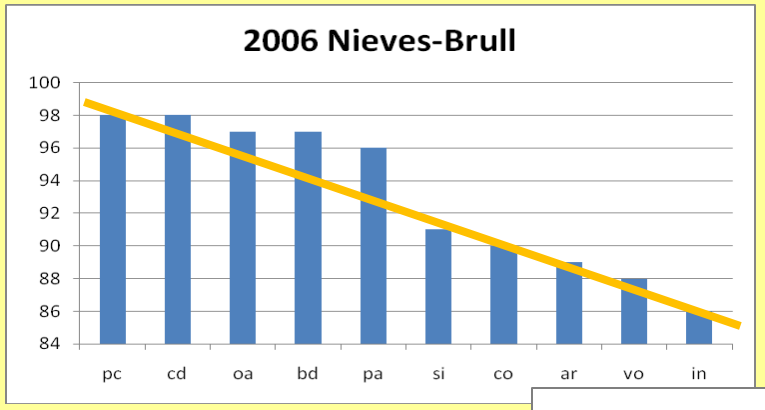
Comparison of mean WISC-R/WISC-III subtest scores

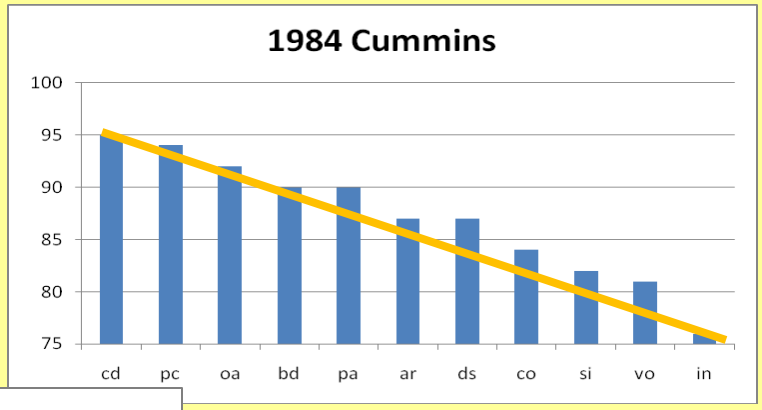
	Hispanic Group (Mercer) (1972)	Hispanic Group (Vukovich & Figueroa) (1982)	ESL Group (Cummins) (1982)	Bilingual Group (Nieves-Brull) (2006)
Subtest Name	**Mean SS**	**Mean SS**	**Mean SS**	**Mean SS**
Information	7.5	7.8	5.1	7.2
Vocabulary	8.0	8.3	6.1	7.5
Similarities	7.6	8.8	6.4	8.2
Comprehension	7.8	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

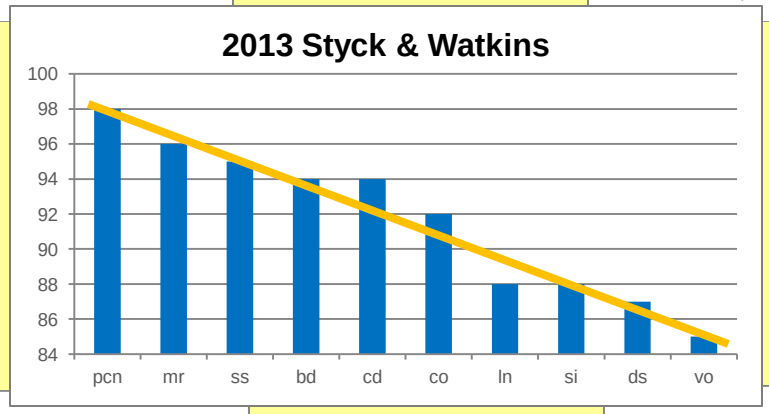
**Data for this subtest were not reported in the study.*

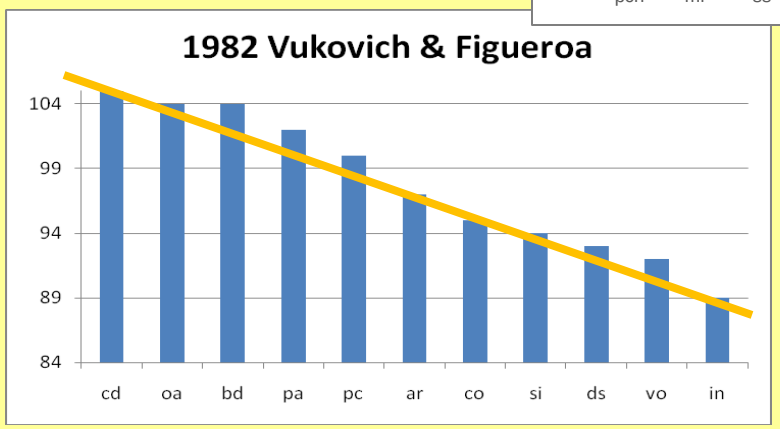
Acculturative Knowledge and Language Proficiency

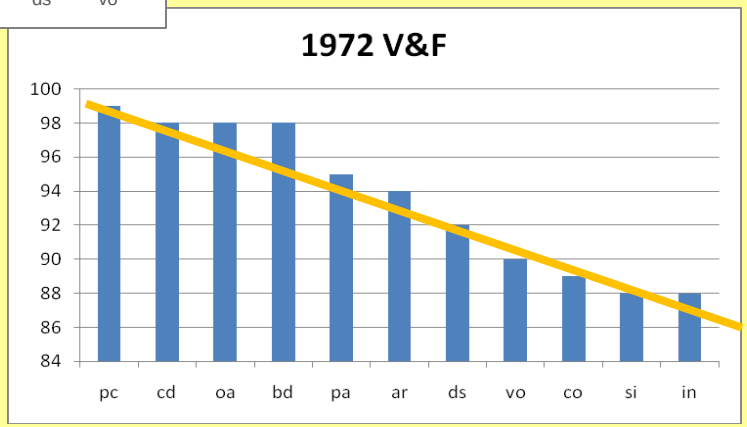
Comparison of WISC-R/WISC-III subtest scores





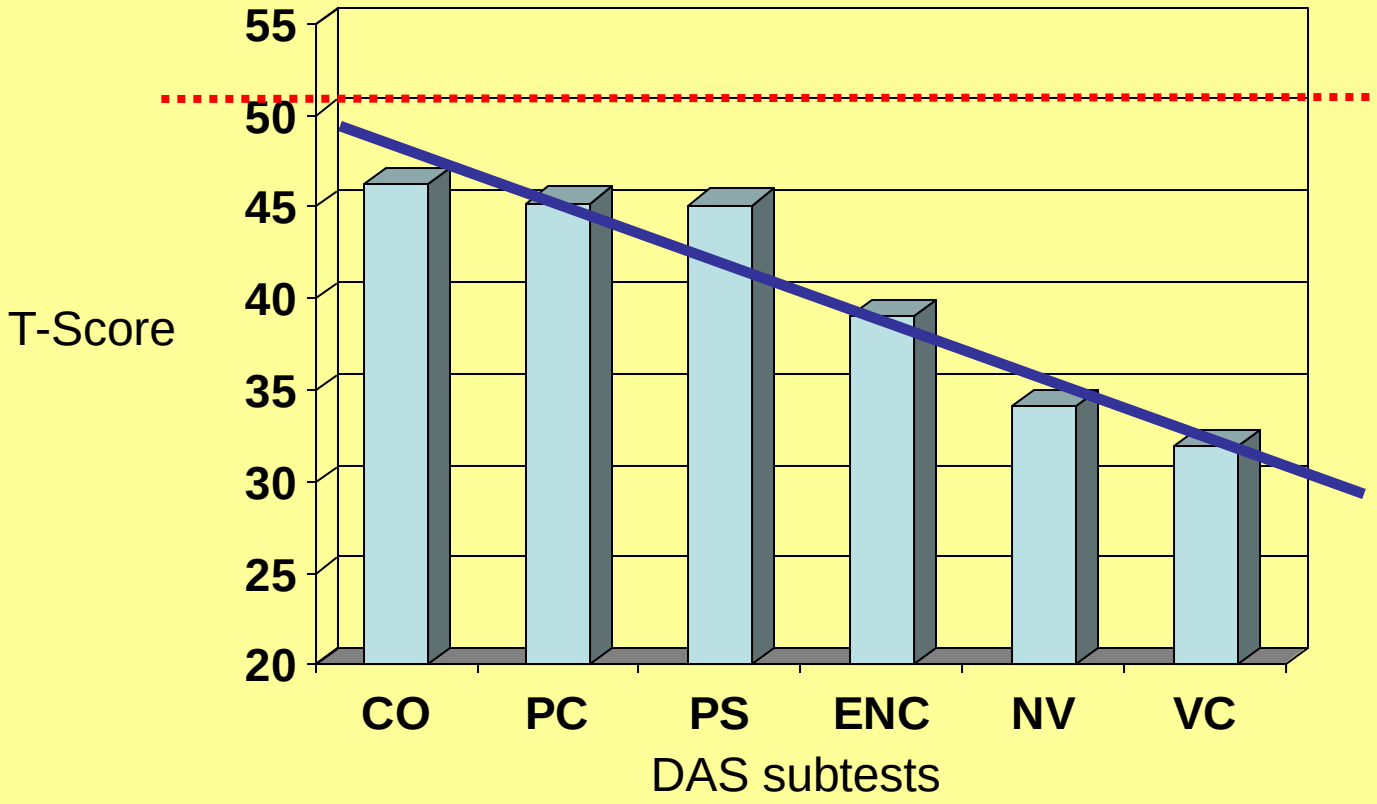






Acculturation and Language Proficiency

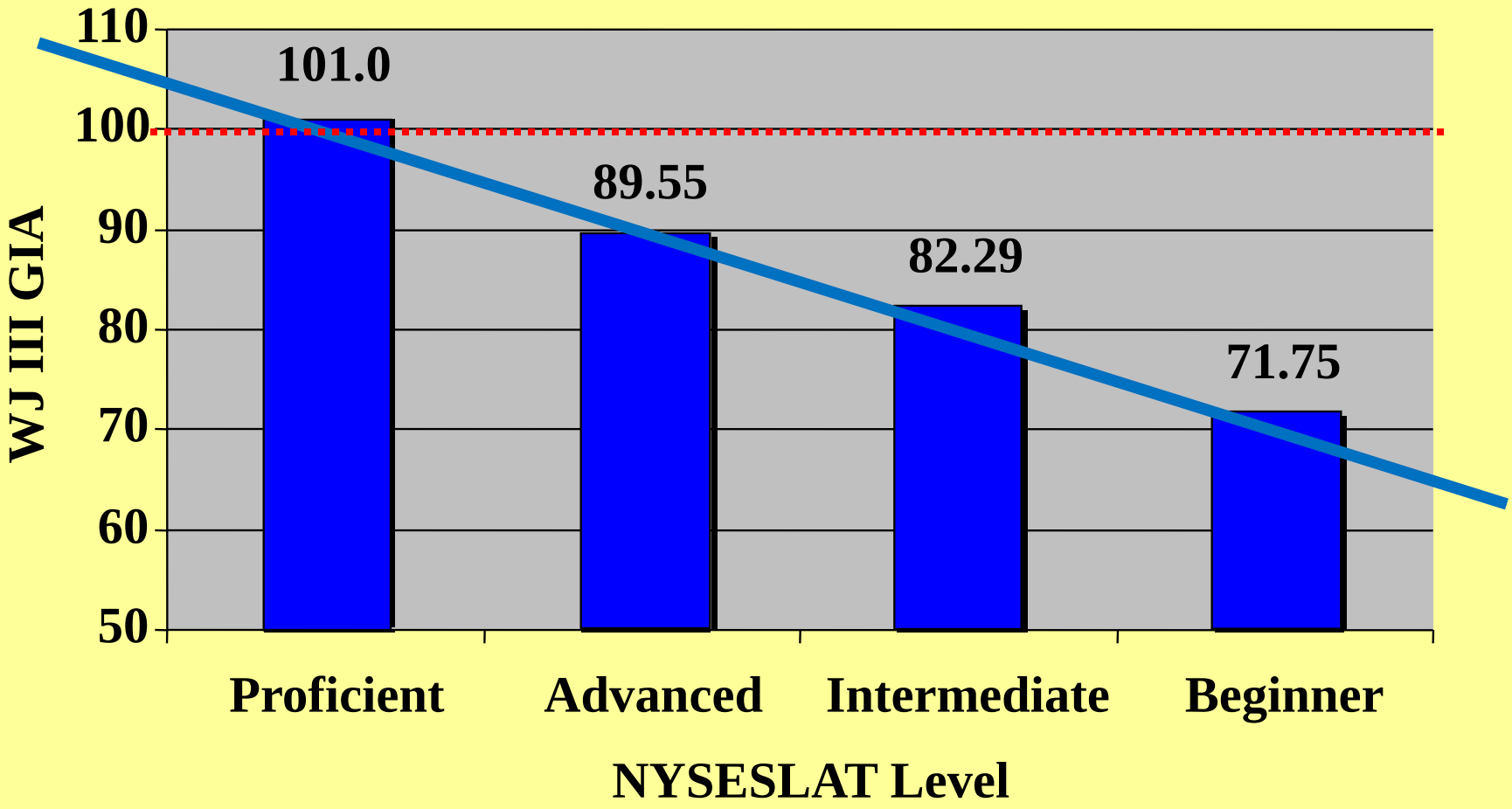
Mean subtest scores across six Differential Ability Scale (DAS) subtests in a pre-school sample of English Language Learners



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

Acculturation and English Language Proficiency

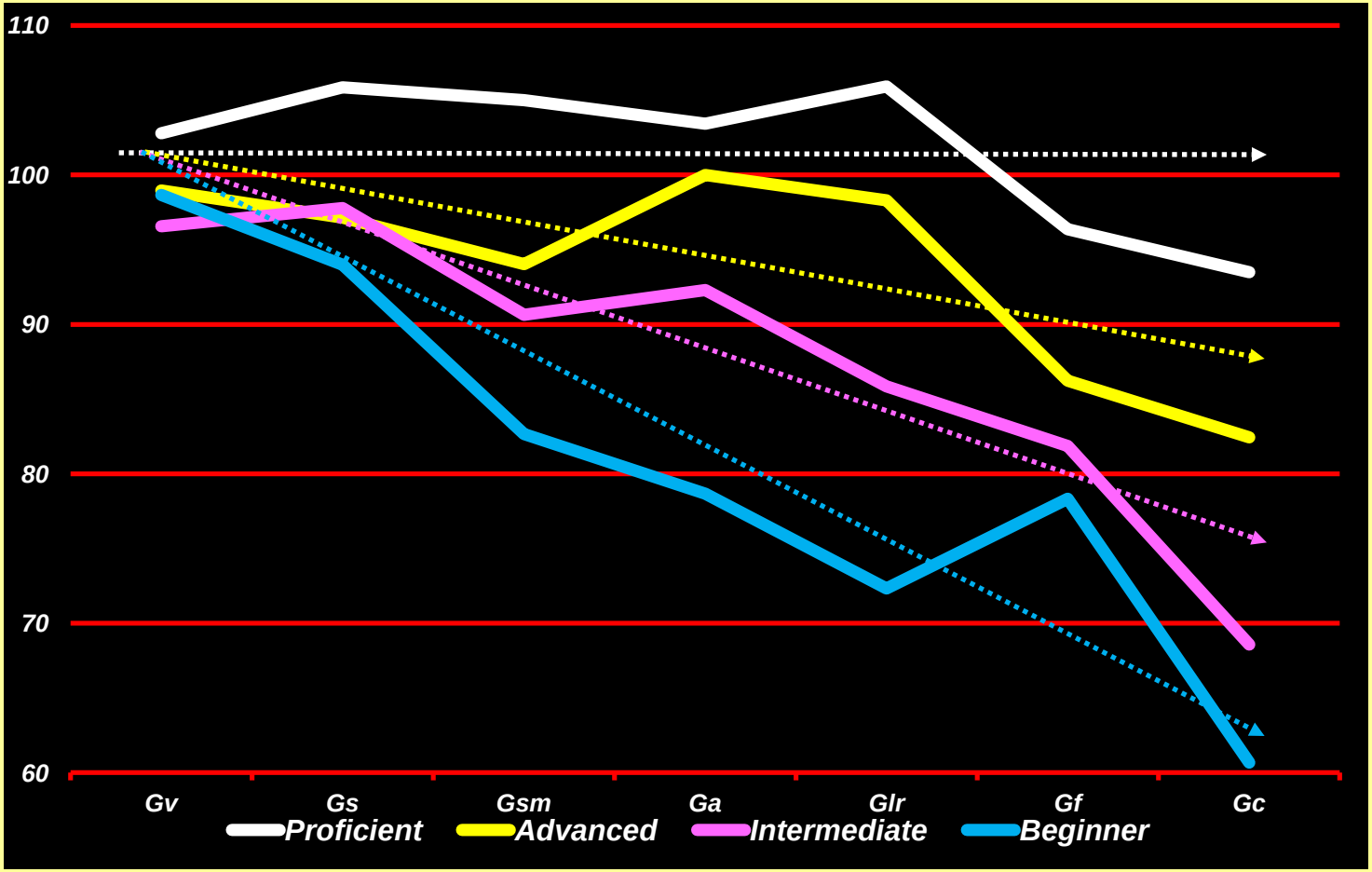
Mean WJ III GIA across the four levels of language proficiency on the New York State ESL Achievement Test



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

Peer-Reviewed Research Done Well: The Empirical Basis of the C-LIM Classifications and Ranges.

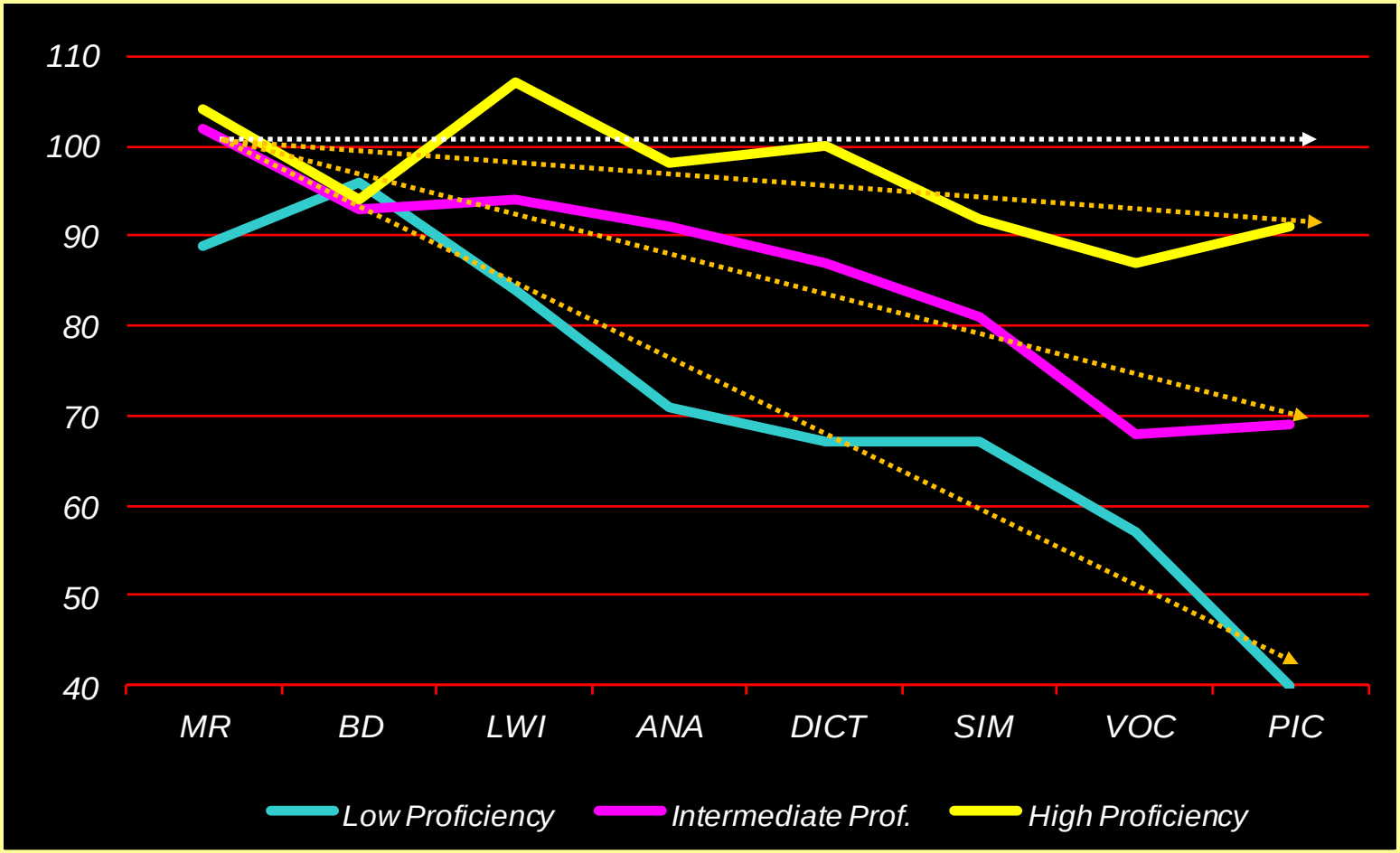
Domain specific scores across the seven WJ III subtests according to language proficiency level on the NYSESLAT



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.

English Language Learner Research Done Well: The Empirical Basis of the C-LIM Classifications and Ranges.

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..

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.)

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

Bibliography of Assessment Resources

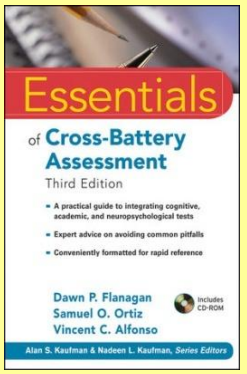
BOOKS:

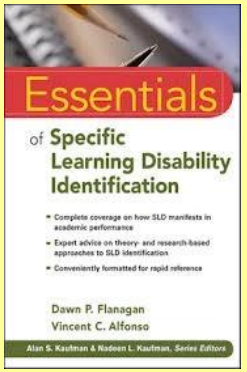
Flanagan, D. P., Ortiz, S.O. & Alfonso, V.C. (2013). Essentials of Cross-Battery Assessment, Third Edition. New York: Wiley & Sons, Inc.

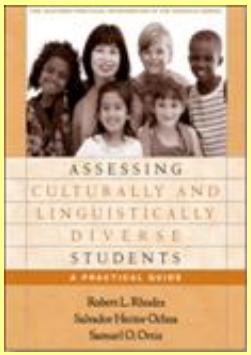
Flanagan, D.P. & Ortiz, S.O. (2012). Essentials of Learning Disability Identification. New York: Wiley & Sons, Inc.

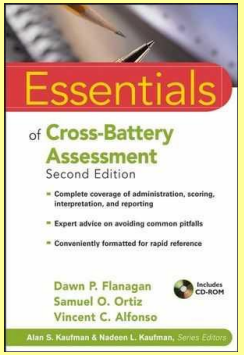
Rhodes, R., Ochoa, S. H. & Ortiz, S. O. (2005). Comprehensive Assessment of Culturally and Linguistically Diverse Students: A practical approach. New York: Guilford.

Flanagan, D. P., Ortiz, S.O. & Alfonso, V.C. (2007). Essentials of Cross-Battery Assessment, Second Edition. New York: Wiley & Sons, Inc.









ONLINE:

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