

Issues in the Assessment of English language Learners

Early Reading First (ERF) FY 2006 New Grantees Meeting

Jamal Abedi

University of California, Davis

National Center for Research on Evaluation, Standards, &
Student Testing, UCLA Graduate School of Education

April 4, 2007

WHY ASSESSMENT IS SO IMPORTANT FOR ELL STUDENTS?

For ELL students assessment starts before instruction

Assessment results affect ELL students in the following areas:

- ***Classification***
- ***Instruction***
- ***Accountability (the NCLB issues)***
- ***Promotion***
- ***Graduation***

Thus assessment of ELL students is very high stakes.

How do ELL students do in assessments in comparison with non-ELL students?

- ELL students perform lower than non-ELL students in general
- The performance-gap between ELL and non-ELL students increases as the language demand of test items increases.
- The performance-gap approaches zero in content areas with a minimal level of linguistic complexity (e.g. math computation)

Site 2 Grade 7 SAT 9 Subsection Scores

Subgroup	Reading	Math	Language	Spelling
LEP Status				
LEP				
Mean	26.3	34.6	32.3	28.5
SD	15.2	15.2	16.6	16.7
N	62,273	64,153	62,559	64,359
Non-LEP				
Mean	51.7	52.0	55.2	51.6
SD	19.5	20.7	20.9	20.0
N	244,847	245,838	243,199	246,818
SES				
Low SES				
Mean	34.3	38.1	38.9	36.3
SD	18.9	17.1	19.8	20.0
N	92,302	94,054	92,221	94,505
High SES				
Mean	48.2	49.4	51.7	47.6
SD	21.8	21.6	22.6	22.0
N	307,931	310,684	306,176	312,321

**Normal Curve Equivalent Means & Standard Deviations
for Students in Grades 10 and 11, Site 3 School District**

	Reading		Science		Math Comp	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Grade 10						
LEP only	24.0	16.4	32.9	15.3	36.8	16.0
LEP & SWD	16.3	11.2	24.8	9.3	23.6	9.8
Non-LEP/SWD	38.0	16.0	42.6	17.2	39.6	16.9
Grade 11						
LEP Only	22.5	16.1	28.4	14.4	45.5	18.2
LEP & SWD	15.5	12.7	26.1	20.1	25.1	13.0
Non-LEP/SWD	38.4	18.3	39.6	18.8	45.2	21.1

**ARE THE STANDARDIZED
ACHIEVEMENT
TESTS APPROPRIATE FOR ELLS?**

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) elaborated on this issue:

For all test takers, any test that employs language is, in part, a measure of their language skills. This is of particular concern for test takers whose first language is not the language of the test. Test use with individuals who have not sufficiently acquired the language of the test may introduce construct-irrelevant components to the testing process. (p. 91)

Are the Standardized Achievement Tests Reliable and Valid for these Students?

- The reliability coefficients of the test scores for ELL students are substantially lower than those for non-ELL students
- ELL students' test outcomes show lower criterion-related validity
- Structural relationships between test components and across measurement domains are lower for ELL students

Site 2 Stanford 9 Sub-scale Reliabilities (Alpha), Grade 9

Sub-scale (Items)	English Only	LEP
Reading, N=	181,202	52,720
-Vocabulary (30)	.835	.666
-Reading Comp (54)	.916	.833
Average Reliability	.876	.750
Math, N=	183,262	54,815
-Total (48)	.898	.802
Language, N=	180,743	52,863
-Mechanics (24)	.803	.686
-Expression (24)	.812	.680
Average Reliability	.813	.683
Science, N=	144,821	40,255
-Total (40)	.805	.597
Social Science, N=	181,078	53,925
-Total (40)	.805	.530

WHY THESE TESTS ARE NOT RELIABLE FOR ELL STUDENTS

**THERE MUST BE ADDITIONAL SOURCES OF
MEASUREMENT ERROR AFFECTING THE
ASSESSMENT OUTCOME FOR THESE STUDENTS**

THESE SOURCES INCLUDE:

- **LINGUISTIC COMPLEXITY OF TEST ITEMS**
- **CULTURAL FACTORS**
- **INTERACTION BETWEEN LINGUISTIC AND
CULTURAL FACTORS WITH OTHER STUDENT
BACKGROUND VARIABLES**

Assumptions of Classical True-Score Test Theory

- 1. $X = T + E$
(Total observed score is the sum of true score plus error score)
- 2. $E(X) = T$
(Expected value of observed score is true score)
- 3. $\rho_{ET} = 0$
(Correlation between error and true scores is zero)
- 4. $\rho_{E_1E_2} = 0$
(Correlation between two error scores is zero)
- 5. $\rho_{E_1T_2} = 0$
(Correlation between error scores and true scores is zero)

Classical Test Theory: Reliability

$$\bullet \sigma^2_X = \sigma^2_T + \sigma^2_E$$

X: Observed Score
T: True Score
E: Error Score

$$\rho_{XX'} = \sigma^2_T / \sigma^2_X$$

$$\rho_{XX'} = 1 - \sigma^2_E / \sigma^2_X$$

Textbook examples of sources of measurement error:

Rater; Occasion; Item; Test Form

How can we improve reliability in the assessment for ELL students?

● **Add more test items**

● **Control for the random sources of measurement error**

● **Control for systematic sources of measurement error**

ADD MORE TEST ITEMS

- “AS A GENERAL RULE, THE MORE ITEMS IN A TEST, THE MORE RELIABLE THE TEST” (SYLVIA & YSSELDYKE, 1998, P. 149).
- “THE FACT THAT A LONGER ASSESSMENT TENDS TO PROVE MORE RELIABLE RESULTS WAS IMPLIED EARLIER...” (LINN, 1995, P. 100).
- “HOWEVER, LONGER TESTS ARE GENERALLY MORE RELIABLE, BECAUSE, UNDER THE ASSUMPTIONS OF CLASSICAL TRUE-SCORE THEORY, AS N INCREASES, TRUE-SCORE VARIANCE INCREASES FASTER THAN ERROR VARIANCE” (ALLEN & YEN, 1979, P. 87).

ADD MORE TEST ITEMS

- **Formula:**

- $$N = \frac{\rho_{xx'}(1 - \rho_{yy'})}{\rho_{yy'}(1 - \rho_{xx'})}$$

- FOR EXAMPLE, IF WE WANT TO INCREASE RELIABILITY OF A 25-ITEM TEST FROM .6 ($R_{YY'}$) TO .8 ($R_{XX'}$), WE NEED TO ADD 43 ITEMS.

A research example showing the effects of increasing test items

• Subscale	N of items	Alpha (α)
• Effort	31	• 0.84
• Effort	17	• 0.90
• Effort	7	• 0.83
• Worry	14	• 0.90
• Worry	11	• 0.90
• Cognitive Strategy	14	• 0.81
• Cognitive Strategy	8	• 0.81

Source: O'Neil, H. F. & Abedi, J. (1996). Reliability and validity of a state metacognitive inventory: Potential for alternative assessment. *Journal of Educational Research*, 89(4), 234-245.

Increasing number of test items for ELL students may cause further complexities:

- **IF THE NEW ITEMS ARE MORE LINGUISTICALLY COMPLEX, THEY MAY ADD TO CONSTRUCT-IRRELEVANT VARIANCE/ MEASUREMENT ERROR AND REDUCE THE VALIDITY AND RELIABILITY EVEN FURTHER**
- **THE COGNITIVE LOAD FOR THE ADDED ITEMS MAYBE GREATER THAN THE COGNITIVE LOAD FOR THE ORIGINAL ITEMS**
- **PROVIDING EXTRA TIME FOR THE NEW ITEMS TO THE ALREADY EXTENDED TIME MAY CAUSE MORE LOGISTICAL PROBLEMS**

Does reliability of a test affect test validity?

- **RELIABILITY SETS THE UPPER LIMIT OF A TEST'S VALIDITY, SO RELIABILITY IS A NECESSARY BUT NOT A SUFFICIENT CONDITION FOR VALID MEASUREMENT (SYLVIA & YSSELDYKE, 1998, P. 177)**
- **RELIABILITY IS A NECESSARY BUT NOT SUFFICIENT CONDITION FOR VALIDITY (LINN, 1995, P. 82)**
- **Reliability limits validity, because $r_{xy} < \sqrt{r_{xx}}$ (Allen & Yen, p. 113)**

FOR EXAMPLE, THE UPPER LIMIT OF VALIDITY COEFFICIENT FOR A TEST WITH A RELIABILITY OF 0.530 IS 0.73

-
- CRESST Studies on the Assessment and Accommodation of ELL Students:

Impact of Language Factors On Assessment of ELLs A Chain of Events

Fourteen studies on the assessment and 3 on the instruction (OTL) of ELL students

Study #1

Analyses of extant data (Abedi, Lord, & Plummer, 1995).

Used existing data from NAEP 1992 assessments in math and science.

SAMPLE: ELL and non-ELLs in grades 4, 8, and 12 main assessment. NAEP test items were grouped into long and short and linguistically complex/less complex items.

Findings

ELL students performed significantly lower on the longer test items.

- ELL students had higher proportions of omitted and/or not-reached items.
- ELL students had higher scores on the linguistically less-complex items.

Study #2

Interview study (Abedi, Lord, & Plummer, 1997)

37 students asked to express their preference between the original NAEP items and the linguistically modified version of these same items. Math test items were modified to reduce the level of linguistic complexity.

Findings

- Over 80% interviewed preferred the linguistically modified items over the original version.

Many students indicated that the language in the revised item was easier:

- "Well, it makes more sense."
- "It explains better."
- "Because that one's more confusing."
- "It seems simpler. You get a clear idea of what they want you to do."

Study #3

Impact of linguistic factors on students' performance (Abedi, Lord, & Plummer, 1997).

Two studies: testing performance and speed.

SAMPLE: 1,031 grade 8 ELL and non-ELL students.
41 classes from 21 southern California schools.

Findings

- ELL students who received a linguistically modified version of the math test items performed significantly better than those receiving the original test items.

Study #4

The impact of different types of accommodations on students with limited English proficiency (Abedi, Lord, & Hofstetter, 1997)

SAMPLE: 1,394 grade 8 students. 56 classes from 27 California schools.

Findings

Spanish translation of NAEP math test.

- Spanish-speakers taking the Spanish translation version performed significantly lower than Spanish-speakers taking the English version.
- We believe that this is due to the impact of language of instruction on assessment.

Linguistic Modification

- Contributed to improved performance on 49% of the items.

Extra Time

- Helped grade 8 ELL students on NAEP math tests.
- Also aided non-ELL students. Limited potential as an assessment accommodation.

Study #5

Impact of selected background variables on students' NAEP math performance (Abedi, Hofstetter, & Lord, 1998).

SAMPLE: 946 grade 8 ELL and non-ELL students. 38 classes from 19 southern California schools.

Findings

- Four different accommodations used (linguistically modified, a glossary only, extra time only, and a glossary plus extra time).
- The glossary plus extra time was the most effective accommodation.

Glossary plus extra time accommodation

- Non-ELLs showed a greater improvement (16%) than the ELLs (13%).
- This is the opposite of what is expected and casts doubt on the validity of this accommodation.

Study #8

Language accommodation for large-scale assessment in science (Abedi, Courtney, & Leon, 2001)

SAMPLE: **1,856** grade 4 and **1,512** grade 8 ELL and non-ELL students.
132 classes from 40 school sites in four cities, three states.

Findings

- Linguistic modification improved performance of ELLs
- No change on the performance of non-ELLs with modified test A published dictionary was both ineffective and administratively difficult as an accommodation
- Different bilingual dictionaries had different entries, different content, and different format.

Psychometric issues in reading assessments

- Reliability and validity issues discussed earlier apply to reading as well
- Reliability and validity coefficients must be clearly examined across the ELL/no-ELL categories
- If a test is reliable for the general population it may not necessarily be reliable for ELLs
- Try to identify and eliminate nuisance variables in the reading assessments

Psychometric issues in reading assessments

We performed several analyses using data reading data from different locations nationwide:

1. Differential functioning in item distractors (DDF) across ELL/non-ELL
2. DDF across SD/non-SD
3. DIF across ELL/nonELL and SD/no-SD

Psychometric issues in reading assessments

- ELL students differentially selected distractors that are longer and/or linguistically complex
- ELL students performed better on items located at the beginning of the test as compared with those located later in the test (item location)
- ELL students had higher level of omit/not-reached items

Psychometric issues in reading assessments for SDs

- There were substantial differences between SDs and non-SDs in their distractor selection
- Non-SDs are selective in their distractor choice
- SDs expressed higher level of fatigue and frustration later in the assessment, more in the second half of the test
- The pattern of DDF was more different across SDs/non-SDs on the second half of the test

Recent publications summarizing findings of our research on the assessment of ELLs :

- Abedi, J. and Gandara, P. (2007). *Performance of English Language Learners as a Subgroup in Large-Scale Assessment: Interaction of Research and Policy*. *Educational Measurement: Issues and Practices*. Vol. 26, Issue 5, pp. 36-46.
- Abedi, J. (in press). *Utilizing accommodations in the assessment of English language learners*. In: *Encyclopedia of Language and Education*. Heidelberg, Germany: Springer Science+ Business Media.
- Abedi, J. (2006). *Psychometric Issues in the ELL Assessment and Special Education Eligibility*. *Teacher's College Record*, Vol. 108, No. 11, 2282-2303.
- Abedi, J. (2006). *Language Issues in Item-Development*. In Downing, S. M. and Haladyna, T. M. *Handbook of Test Development* (Ed.). New Jersey: Lawrence Erlbaum Associates, Publishers.
- Abedi, J. (in press). *English Language Learners with Disabilities*. In Cahlan, C. & Cook, L. *Accommodating student with disabilities on state assessments: What works?* (Ed.) New Jersey: Educational Testing Service.
- Abedi, J. (2005). *Assessment: Issue and Consequences for English Language Learners*. In Herman, J. L. and Haertel, E. H. *Uses and Misuses of Data in Accountability Testing* (Ed.) Massachusetts: Blackwell Publishing Malden.

Conclusions and Recommendation

Assessment for ELL students:

- Must be based on a sound psychometric principles
- Do not assume that psychometric characteristics of assessment are the same across the subgroups
- Must be controlled for all sources of nuisance or confounding variables
- Must be free of unnecessary linguistic complexities
- Must include sufficient number of ELLs in its development process (field testing, standard setting, etc.)
- Must be free of biases, such as cultural biases