Status 2007:

Inspecting the Validity of Large-Scale Assessment Score Inferences for ELLs and Others under More Optimal Testing Conditions—Does it Measure Up?

Rebecca J. Kopriva

David E. Wiley

Jessica Emick

Paper commissioned for the Assessment and Accountability Comprehensive Center, WESTED, San Francisco, CA. April 2007 and adapted from the presentation at the American Educational Research Association Annual Meeting, April 11, 2007.

Abstract

The goal of the current study was to examine the influence of providing more optimal testing conditions and evaluate the effect this has the validity of the score inferences across ELL students with different needs, strengths, and levels of language proficiency. It was expected that the validity of the score inferences would be similar for 3rd and 5th grade ELL students with different needs and exited and native English speakers who acted as control groups. Multiple choice and constructed response mathematics data from a large-scale data collection were analyzed relative to data from a criterion measure developed for the study and other ancillary information obtained during the project. Results indicated the validity data from multiple choice results for ELLs were generally very poor compared to the control groups, but that validity data from constructed response scores were more promising, especially for beginning students in both grades and advanced students in grade 5. Additional analyses indicated significantly higher misclassification rates of test score data for lower English proficient ELL students as compared to control groups when looking at students classified as knowing some mathematics on the criterion measure. This study raises many questions about the validity of inferences drawn from large-scale assessments for students with lower English proficiency. It also calls into question the effectiveness of measuring content knowledge with some item types for students at various levels of English language proficiency, and suggests that item type may interact with grade level of the test takers.

Research Question

This study was designed to provide appropriate test accommodations to students who needed them in a large-scale testing setting. Needs of individual students were identified for two populations of test takers: ELLs and native English speaking students who were classified as poor readers. Mathematics test items were developed to provide enhanced access to students with language difficulties, bilingual and picture-word glossaries for selected ELLs were created, and various administration accommodations designed to support the test materials were used. Individualized screening to match students and accommodations was undertaken to ensure that most students received what they needed. The items were administered as part of a district-wide large-scale assessment, and implementation of accommodations was supported by project staff and monitored for quality control. Because of these steps, the hypothesis to be tested was that, for students with different needs and across those who received different sets of accommodations, the relationship between test scores and a criterion measure of student mathematics knowledge and skills would be similar for ELLs and the control groups (exited ELLs and native English speakers).

Method

This study is part of a larger project, The Valid Assessment of English Language Learners (VAELL, Kopriva & Mislevy, 2005). The data for this study were collected in the fall and winter of 2004/2005.[1]

1. Sample

Data were obtained from 2502 third- and fifth-grade students from 21 schools in a school district in Maryland. 19 of the schools were selected as those who had high levels of English learners. An additional two schools were identified where student scores on previous tests were the highest in the district. These schools were added to ensure that the range of scores and abilities were included in the study. All 3rd and 5th grade classes from these schools participated. In total, several hundred ELL students, were identified, who varied in language of origin, language acquisition status, other language development variables (e.g., length of time in country, level of reading, writing and mathematics achievement, and demographic characteristics). Native speakers who were poor readers were also identified, and exited ELLs and non-ELLs acted as control groups. Table 1 breaks down students by ELL group (including exited and non-ELL students).

[1] Several researchers worked on the project, designing and implementing the data collection used in this study, and contributing to the original data cleaning and analyses of the project data. These included Chen Su Chen, Carol Boston, Amy Henderson, Jessica Emick, Cathy Cameron, Heather Mann, Peng Lin, and Bob Mislevy. Their work was instrumental in ensuring the high quality of the data used here.

Table 1: Frequency (%) of Students in Each ELL Level		
	Grade 3	Grade 5
Beginning	52 (4.1%)	46 (3.7%)
Intermediate	198 (15.4%)	148 (12.1%)
Advanced	75 (5.9%)	55 (4.5%)
Exited	245 (19.1%)	256 (20.9%)
Non-ELL	711 (55.5%)	719 (58.8%)
	N = 1281	N = 1224

2. Instruments and Additional Accommodations

Prior to the administration of the mathematics assessment, teachers of participating students completed a questionnaire for each student concerning the student's mathematics abilities. The questionnaire also collected data on ancillary student characteristics that might impact student performance. For the mathematics section, the questionnaire asked teachers to rate (on a three-point scale: rarely, sometimes, almost always) how often the students successfully demonstrated knowledge and skills of particular mathematics construct elements in the classroom. The abilities that were targeted were what were being measured by the 19 items. For instance, for third-graders, one question asked teachers to rate prevalence of classroom performance on the following element: *This student can solve a word problem involving a solution requiring subtraction with regrouping*. About 20 questions were asked to 3rd and 5th grade teachers about each student's abilities, respectfully.

These ratings would become the criterion indicator of student achievement, and were used in lieu of a standardized test score because of the confounding problems of language and target abilities in most tests. The rating system used was at a similar level

of detail to one used by Schmidt, McKnight, Houang, Wang, Wiley, Cogan and Wolfe et al. (2001) in their analyses of TIMSS text book elements and curriculum data across countries. These researchers have since used this approach in other studies where it has been found to be replicable indicator for making differential judgments about content. It was also consistent with how the state of Maryland and the district identified specific instructional objectives in their content standards, which meant that the participating teachers were familiar with this approach to considering aspects of the curriculum. For this investigation, the approach was initially piloted and then refined to be useable and feasible for teachers to differentiate student ability.

Besides demographic information, the ancillary data collected on the questionnaire included use of strategies in mathematics problem solving, assessment experiences in the classroom and on other large-scale tests, English language arts skills, learning strengths and challenges, and factors that are hypothesized to either support or inhibit student access in testing math content.

Subsequently**,** students participated in the district benchmark mathematics test which was developed to mirror the state's large-scale assessment. The benchmark test was developed by district officials in an effort to provide students a school-wide practice trial prior to the official testing window. In all, 11 multiple choice items and 8 constructed-response items were inserted in the district's assessment at each grade. These items were keyed to the state's standards and indicators, tied to the district plan as curriculum already taught in the district's third and fifth grade classes in the current academic year, and approved by the district as measuring content that the district would have otherwise written items to reflect. The 19 items were rewritten versions of

mathematics items released from other states. These versions were designed to measure the same mathematics constructs but provide more access for students with less proficiency (e.g., shorter sentences, modified vocabulary, more accessible problem contexts, clearer formatting, use of pictures/graphic organizers, etc.). Mathematics experts had previously reviewed the items to ensure that they were measuring the same targeted mathematics knowledge and skills as the original items.

3. Procedures

After completed teacher questionnaires for each student were received by study staff, accommodations were assigned based on needs and challenges of individual students as identified in the teacher questionnaire. The algorithms that were used to match individual students to specific accommodations were an early prototype upon which a later product, the empirically supported *Selection Taxonomy for English Language Learners Accommodations* (*STELLA*), was created (for instance see Koran and Kopriva, in press, for an explanation of *STELLA*). Students could receive no accommodations, or one or more accommodations in a package. All students were assigned the accommodations deemed essential to their ability to access the test, within the logistical district constraints and constraints of the scope of accommodations used in the study.

For the purposes of this study, identified ELLs and some poor readers received individualized sets of accommodations. Accommodations that were variously assigned included

- Spanish-English glossary
- word-picture list in English
- use of manipulatives

- oral administration in English
- small-group administration
- access to a bilingual language liaison.

To implement the administration of the accommodations, staff were hired and trained. Because of the short administration time window (a week total), sufficient staff to concurrently cover several schools and several classes within a school at the same time meant that a large volume of qualified staff were needed. Publicity efforts were launched in the fall and participants were recruited through fliers around campus, through campus and public newspapers, and to community organizations. Participants were subsequently screened and selected to take part in the training. Training occurred within the month preceding the assessment administration, and, during the training, participants went through a second screening. In all, several hundred staff were finally hired to take part in the study and implement the accommodations.

The administration of the mathematics test and its language arts counterpart occurred over two to three days at each school, depending on school arrangements. All students were administered the mathematics benchmark test, including the 19 items identified for this study. All large-scale standard administration procedures were followed for most students. For students who were identified as needing accommodations, typical large-scale administration procedures of the accommodations were used. For the accommodations which were part of this study, administration was monitored by study staff.

Results

1. Descriptive

Table 2 presents the mean and standard deviations of the test score results by grade, group, and item type and Table 3 provide these data for the teacher ratings. As expected, for both grades the mean test scores are higher for exited ELLs and non-ELLs than for the three ELL groups, and within ELLs, variability in the test scores increase as students gain more English proficiency. Of interest, the scores of the exited students in grade 3 were higher than non-ELLs for both multiple choice and constructed response subtests, and higher for the constructed response subtest for grade 5.

As illustrated in Table 3, the criterion measure mirrors the increase in average ratings as language proficiency increases. However, across groups the variability in the target criterion ratings remains largely consistent for all levels of proficiency. The variability suggests that teachers of students at all levels of English appear to be able to differentiate the students' mathematics knowledge; the consistency of the variability across groups suggests that teachers were able to differentiate across the same range of ability for ELLs as well as for natives and exited.

Table 2: Mean and Standard Deviation of Mathematics Test Scores by ELL Level				
	Grade 3		Grade 5	
	Multiple Choice	Constructed Response	Multiple Choice	Constructed Response
Beginning	3.192 (1.401)	2.154 (1.775)	3.022 (1.485)	2.244 (1.694)
Intermediate	3.832 (1.919)	2.919 (2.459)	3.757 (1.940)	4.236 (3.290)
Advanced	5.293 (2.235)	4.253 (2.853)	4.800 (2.305)	5.564 (3.553)
Exited	6.318 (2.609)	6.392 (3.524)	5.800 (2.382)	7.765 (3.919)
Non-ELL	5.899 (2.554)	5.752 (3.748)	5.809 (2.705)	6.752 (4.169)
	N = 1281		N = 1224	

Table 3: Mean and Standard Deviation of Criterion Mathematics Rating by Teachers by ELL Level				
	Grade 3		Grade 5	
	Multiple Choice	Constructed Response	Multiple Choice	Constructed Response
Beginning	1.610 (0.520)	1.590 (0.576)	1.380 (0.477)	1.304 (0.453)
Intermediate	1.852 (0.475)	1.905 (0.514)	1.619 (0.520)	1.531 (0.513)
Advanced	2.210 (0.431)	2.256 (0.462)	1.934 (0.520)	1.837 (0.549)
Exited	2.391 (0.428)	2.443 (0.431)	2.249 (0.559)	2.167 (0.584)
Non-ELL	2.291 (0.499)	2.332 (0.509)	2.214 (0.572)	2.160 (0.613)
	N = 1281		N = 1224	

2. Regressions and post hoc Comparisons

These analysis looked at the relationship between the criterion measure and the test score to investigate if the relationships were similar for all groups—particularly if the relationship was similar for levels of ELL and native speakers, and ELLs and exited students. We were interested in both the ability of the test scores to differentiate student ability (as defined by the target ratings), and the amount of predictive variation in the relationship. For this analysis, the test score was the dependent variable and the criterion rating was the independent variable.

While the R or R_{xy}^2 is the typical coefficient which researchers use to estimate the relationship between two variables, we believe that both this indicator and the coefficient of the target criterion (the beta or β) provide meaningful information. We expect a reasonable R square with a reasonably large beta for each group and that these results will be similar across groups. As Equation 1 illustrates, R_{xy}^2 is actually a composite indicator of both the slope of the relationship and the variation (σ_θ). A reasonable R square could be the result of a reasonable β with relatively little variation around the line. However, a smaller R_{xy}^2 could also include a reasonable beta if the variance (σ_θ) around the line is large. Conversely, a larger R_{xy}^2 could occur if both the B and the σ_θ are small, and a smaller R_{xy}^2 would also occur if the β is small and the σ_θ is large. While producing a larger R square, it does not seem that a small β and small variation would be very indicative of a useful relationship.

Equation 1 $R_{xy} = Cov(x,y)/((SD(x))(SD(y)))$

$= \beta\sigma_x^2/(\sigma_x)(\sqrt{(\beta^2 \sigma_x^2 + \sigma_\theta^2)})$

$= \beta \sigma_x/\sqrt{(\beta^2 \sigma_x^2 + \sigma_\theta^2)}$

$R_{xy}^2 = \beta^2 \sigma_x^2/(\beta^2 \sigma_x^2 + \sigma_\theta^2)$

a. Grade 3 Regressions

Regression results for Grade 3 are presented in Table 4, and contrasts of each set of raw betas are reported in Table 5. The significance tests for the contrasts were completed by performing Analysis of Covariance analyses where the dependent variable was test score, the independent variable was the contrasted groups, and the covariate was the teacher ratings. The F-ratio that is presented is the interaction F. This analysis answers the question: Did the covariate interact differently for the first group as compared to the second?

As the findings in Table 4 indicate, the R^2 and betas differ substantially across groups for both multiple choice and constructed response scores. Specifically, the table reports that R squared relationships are much larger between the two measures for the Exited and non-ELLs as compared to their ELL peers in most cases. It also illustrates that, for the multiple choice items, the beta is not even significantly different from 0 for either the beginner or advanced students. Table 5 confirms that the beta coefficients are generally not equivalent as most of the ELL betas are significantly different than the betas for either exited or non-ELLs. The one exception to this is the beta contrast for beginners vs. non-ELLs for the constructed response subtest where the result is not

significant. This difference may have been overly effected by the unequal sample sizes (and SE's), but since the R^2 is also high, the result is promising.

Table 4. Grade 3 Regression Results

		IV	B	S.E.	P	R^2	F	p
Multiple Choice	Beginning	(Constant)	2.538	0.636	0.000	0.023	10.400	0.000
		Target	0.407	0.376	0.285			
	Intermediate	(Constant)	2.228	0.539	0.000	0.046		
		Target	0.866	0.282	0.002			
	Advanced	(Constant)	2.838	1.334	0.037	0.046		
		Target	1.111	0.593	0.065			
	Exited	(Constant)	-0.208	0.849	0.807	0.200		
		Target	2.728	0.349	0.000			
	Non-ELL	(Constant)	-0.220	0.384	0.567	0.273		
		Target	2.670	0.164	0.000			
Constructed Response	Beginning	(Constant)	-0.094	0.656	0.886	0.209	8.670	0.000
		Target	1.413	0.388	0.001			
	Intermediate	(Constant)	0.723	0.657	0.272	0.058		
		Target	1.153	0.333	0.001			
	Advanced	(Constant)	0.331	1.598	0.836	0.079		
		Target	1.739	0.694	0.015			
	Exited	(Constant)	-2.592	1.162	0.027	0.202		
		Target	3.679	0.469	0.000			
	Non-ELL	(Constant)	-2.565	0.578	0.000	0.234		
		Target	3.568	0.242	0.000			

Table 5. Grade 3 Contrasts of Slopes

		Multiple Choice			*Constructed Response*		
Contrast	Df	F-ratio	p-value	sign	F-ratio	p-value	sign
1 vs 2	1,246	1.000	0.318		.097	0.756	
1 vs 3	1,123	.525	0.470		.162	0.688	
1 vs 4	1,293	11.101	0.001	*	3.972	0.047	*
1 vs 5	1,761	13.436	0.000	*	3.582	0.059	
2 vs 3	1,269	.002	0.969		.038	0.846	
2 vs 4	1,439	14.537	0.000	*	13.000	0.000	*
2 vs 5	1,907	21.851	0.000	*	14.832	0.000	*
3 vs 4	1,316	5.982	0.015	*	6.258	0.013	*
3 vs 5	1,784	7.561	0.006	*	6.269	0.013	*
4 vs 5	1,954	.012	0.919		.009	0.923	

b. Grade 5 Regressions

For Grade 5, results are presented in Table 6 and Table 7. As the findings in Table 6 indicate the equivalence of the R^2 and betas for this grade generally show more potential. For both item types, the intermediate students did not reach parity with either of control groups. And, as in grade 3, the beta for the beginner group in the multiple choice is not significantly different from 0 which, along with the R square, indicates a non-existent relationship. However, in constructed response, the betas for beginners are not significantly different form either non-ELLs or exited, and the R^2 is promising for beginners (although it is still different for beginners as compared to the control groups). Further, the R^2 for the advanced ELL group for the multiple choice subtest is close to exited, and it is close to non-ELLs in constructed response. Table 7 illustrates that the betas for the advanced group are not significantly different from either exited or non-ELL groups, and this finding holds over both multiple choice and constructed response.

These findings suggest that for advanced ELLs, they appear to reach parity with non-ELLs and exited students in both their multiple choice and constructed response. Even though the R squares for the multiple choice regressions are somewhat different, the strength of the slope is consistent (albeit with more variation around the line for advanced than for the control groups). Likewise, for beginner ELLs on constructed response, the strength of the relationship is consistent (as reflected by the equivalent B's) even though the variation in the relationship is greater for beginners as compared to exited and non-ELLs. Of note is the broad distinction between the findings for the beginners on the multiple choice versus the constructed response scores and the sobering findings for intermediate ELLs on both item types.

Table 6. Grade 5 Regression Results								
		IV	B	S.E.	p	R^2	F	p
Multiple Choice	Beginning	(Constant)	3.461	0.688	0.000	0.011	6.410	0.000
		Target	-0.318	0.472	0.504			
	Intermediate	(Constant)	2.497	0.514	0.000	0.044		
		Target	0.779	0.302	0.011			
	Advanced	(Constant)	1.684	1.135	0.144	0.132		
		Target	1.611	0.567	0.006			
	Exited	(Constant)	1.594	0.558	0.005	0.193		
		Target	1.870	0.241	0.000			
	Non-ELL	(Constant)	0.828	0.355	0.020	0.227		
		Target	2.249	0.155	0.000			
Constructed Response	Beginning	(Constant)	0.521	0.736	0.483	0.125	2.650	0.032
		Target	1.322	0.534	0.017			
	Intermediate	(Constant)	1.818	0.831	0.030	0.061		
		Target	1.580	0.515	0.003			
	Advanced	(Constant)	0.215	1.519	0.888	0.203		
		Target	2.910	0.793	0.001			
	Exited	(Constant)	0.325	0.814	0.690	0.262		
		Target	3.433	0.363	0.000			
	Non-ELL	(Constant)	-0.142	0.504	0.778	0.220		
		Target	3.192	0.224	0.000			

Table 7. Grade 5 Contrasts of Slopes							
		Multiple Choice			*Constructed Response*		
Contrast	df	F-ratio	p-value	sign	F-ratio	p-value	sign
1 vs 2	1,190	3.062	0.082		0.086	0.770	
1 vs 3	1,97	6.442	0.013	*	2.598	0.110	
1 vs 4	1,298	9.605	0.002	*	3.656	0.057	
1 vs 5	1,762	11.366	0.009	*	2.318	0.128	
2 vs 3	1,199	1.971	0.162		2.271	0.133	
2 vs 4	1,400	6.790	0.010	*	8.183	0.005	*
2 vs 5	1,864	13.110	0.000	*	6.267	0.013	*
3 vs 4	1,307	.123	0.726		0.182	0.670	
3 vs 5	1,771	.939	0.333		0.014	0.905	
4 vs 5	1,972	1.785	0.182		0.401	0.527	

3. Comparison of Probability Distributions

Because of the generally discouraging results for grade 3 and the more mixed but somewhat promising results for grade 5, the decision was made to further inspect the relationships between the criterion and the subtest scores. In particular, we were interested in understanding how the relationships might differ between groups, for students that scored very low (below chance or a similar level in constructed response) and higher on the test and yet their teachers said they knew the material at least some extent. Two sets of chi-square analyses were conducted. The first inspected the quadrant distributions (of test scores x criterion ratings) of each of the ELL groups (beginner, intermediate, advanced) and exited versus the native English speakers. The second used the same quadrant distributions and examined how particular ancillary variables impacted control and treatment groups for each of the quadrants. While results won't change appreciatively, the decision was also made to further narrow the non-ELL group and take out the few ELL students whose parents chose to keep out of ELL services. Therefore, for grade 3, native speakers = 675 (as compared to non-ELL=711) for a difference of 36. For grade 5 the difference is 49 students (n=670 as compared to 719).

Since the focus of the inspection was to view the test score/rating relationship for students who teachers said knew some mathematics, but where their score did not reflect this knowledge, the teacher bar was purposely set higher than the score bar. The score bar was purposely set very conservatively—chance levels for the multiple choice portions and ¼ of the total possible scores for the constructed response scores. The teacher judgment criterion bar on the other hand was set at 1/3 of the total possible ratings. This was done so that the intended quadrant (students whose teachers said had mathematics

ability but who had received chance level scores) would underestimate the students who knew some mathematics and thereby increase the probability that the quadrant was reflecting the performance of students with true score mathematics ability.

We reasoned that by using these cut points, fewer possible classifications in the LH quadrant classification (low on scores, high on ratings) would turn out to be false negatives. In other words, if significant differences between the quadrants for the ELL levels vs. native speakers were found, this finding would tend to underestimate the problem rather than over estimate it. (We are aware that by making these decisions, we are over representing students who may be incorrectly categorized in the low rating group—both those who will be consistently classified using both measures (LL) or inconsistently classified (HL)).

Table 8 presents the cut scores for both indices. Teacher ratings ranged from 1 to 3 per item for a total of X-Y points. Teacher rating averages per subtest were used. Mathematics subtest scores ranged from 0 to 11 for multiple choice, both grades, and 0-15 (grade 3) and 0-17 (grade 5) for constructed response. Total points within each subtest were used for this study.

Table 8. Cut Points for Quadrants

	Grade 3	
	Measure	Cut
Multiple Choice	TR Scores	<1.7 <3
Constructed Response	TR Scores	<1.7 <4
	Grade 5	
	Measure	Cut
Multiple Choice	TR Scores	<1.7 <3
Constructed Response	TR Scores	<1.7 <4

The hypotheses for the first set of analyses are that, for 3rd grade, all levels of ELL students will have significantly higher rates of LH misclassification than will exited and native speakers. We expect that 5th grade beginner and intermediate ELL students will also have significantly higher rates of LH misclassification than will advanced, exited and native speakers. For the second set of analyses, even though every effort was made to minimize the impact of the ancillary variables on the test scores, the hypotheses are that ancillary variables will continue to impact the scores of the aforementioned ELL students from their respective grades more than the exited or native speakers (and advanced ELLs for grade 5). We hypothesize that the increased impact of the variables will lead to significantly higher rates of possible misclassification for those ELL students.

a. Comparison of Quadrant Distributions for ELL groups/Exited vs. Natives

Grade 3 Distribution Results Figures 1 through 10 illustrate the distribution results graphically for each group. Figures 1-5 utilize the multiple choice test results and teacher ratings and 6-10 use the constructed response test results/teacher ratings. The teacher ratings are displayed on the horizontal axis, test scores are on the vertical axis. The vertical and horizontal lines inside the graphs designate the quadrants by indicating the cutpoints for both the test scores and teacher ratings, respectively, while the diagonal line illustrates the slope of the solution. In the graphs, dots are at each score point received by one or more students, and in most cases the dots reflect scores/ratings for multiple students. In each graph the lower left (low test score, low teacher rating (LL)) and upper right (high test score, high teacher rating (HH)) indicate consistent classifications for the two measures, while the other quadrants reflect inconsistent classifications. As noted above, the primary quadrant of interest is the LH (low test scores

and high teacher rating) and this quadrant is found on the bottom right hand side of each graph.

Overall, the graphs seem to suggest that ELL students with lower English proficiency may have a greater percentage of students who are misclassified as LH than do exited, native speakers, and sometimes more proficient ELL students. This is true for both multiple choice and constructed response.

Insert Figures 1-10

Table 9 presents the grade 3 frequency and percentage distribution data by quadrant for multiple choice and constructed response results in each of the 5 groups. It also gives the targeted conditional probability data of inconsistently classified students—the subset of low scoring students among those the teachers rated as having some mathematics knowledge (L/H). A significant chi-square for the total sample (X^2 =159.9, df=12, p<.0001) indicates the multiple choice distributions over quadrants fluctuate among the groups. For constructed response results, the chi-square results are also highly significant (X^2 = 213.5, df=12, p.<0001). For both the multiple choice and constructed response results, substantially greater percentages of ELLs (beginner and intermediate in particular for the multiple choice, and all levels for constructed response) vs. the control groups were misclassified as L given criterion ratings as H.

Table 9. Total Sample Distributions by Quadrant, Grade 3

	Group	HH	HL	LH	LL	P(L/H)
	Beginner	17(.33)	19(.37)	7(.14)	9(.17)	.29
Multiple	Intermediate	104(.53)	39(.20)	33(.17)	22(.11)	.24
Choice	Advanced	59(.79)	8(.11)	8(.11)	0(.00)	.12
	Exited	214(.87)	9(.04)	19(.08)	3(.01)	.08
	Native	544(.81)	63(.09)	47(.07)	21(.03)	.08
	Beginner	7(.14)	4(.08)	8(.15)	33(.64)	.53
Constructed	Intermediate	37(.19)	31(.16)	58(.29)	72(.36)	.61
Response	Advanced	34(.45)	5(.07)	24(.32)	12(.16)	.41
	Exited	169(.69)	16(.07)	38(.16)	22(.09)	.18
	Native	383(.57)	50(.07)	148(.22)	94(.14)	.28

To test the degree of difference in the L/H conditional probabilities, Table 10 reports the results from the critical ratio with chi square test of significance for each of the ELL groups and exited vs. native speakers. As expected, the beginner and intermediate are significantly different than native speakers for both multiple choice and constructed response results. The advanced is significantly different than the native for constructed response, but not for multiple choice. A similar P(L/H) in the multiple choice subtest for advanced suggests that the lower regression findings for this group stem from general variation which does not include inconsistent classification of those students teachers rated as having at least some ability in mathematics.

Table 10. Grade 3 Total Sample Conditional Probability P(L/H) Comparisons

		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Beg vs Nat	0.212	0.093	2.270	5.154	1	0.023
	Int vs Nat	0.161	0.038	4.225	17.849	1	0.000
	Adv vs Nat	0.040	0.041	0.969	0.939	1	0.332
	Exit vs Nat	0.002	0.021	0.096	0.009	1	0.924
Constructed Response	Beg vs Nat	0.255	0.130	1.954	3.820	1	0.051
	Int vs Nat	0.332	0.054	6.181	38.207	1	0.000
	Adv vs Nat	0.135	0.068	2.000	4.000	1	0.045
	Exit vs Nat	-0.095	0.033	-2.865	8.210	1	0.004
N = 1245							

Grade5 Distribution Results

Figures 11 through 20 illustrate the distribution results graphically for each group in grade 5. Like grade 3, the graphs for both multiple choice and constructed response results seem to suggest that ELL students with lower English proficiency may have a greater percentage of students who are misclassified as LH than do exited and native speakers.

Insert Figures 11-20

Table 11 presents the grade 5 conditional probability and frequency and percentage distribution data by quadrant for multiple choice and constructed results. Like grade 3, very significant chi-squares for both multiple choice and constructed response data (X^2 =169.7, df=12, p<.0001 and X^2 = 211.9, df=12, p.<0001, respectively) indicates the distributions over quadrants fluctuate among the groups. For multiple choice results, large differences in the L/H probabilities between beginner and exited/native are evident, as are the constructed response differences between beginner and intermediate groups vs. exited/native. Substantial differences also seem to occur for intermediate and exited/native groups in multiple choice results, and for advanced and native vs. exited in constructed response.

Table 11. Total Sample Distributions by Quadrant, Grade 5

	Group	HH	HL	LH	LL	P(L/H)
	Beginner	5(.11)	24(.52)	6(.13)	11(.24)	.55
Multiple	Intermediate	49(.33)	56(.38)	14(.10)	29(.20)	.22
Choice	Advanced	33(.60)	14(.26)	5(.09)	3(.06)	.13
	Exited	195(.76)	41(.16)	13(.05)	7(.03)	.06
	Native	482(.72)	106(.16)	43(.06)	39(.06)	.08
	Beginner	3(.07)	6(.13)	5(.11)	32(.70)	.63
Constructed	Intermediate	29(.20)	47(.32)	25(.17)	47(.32)	.46
Response	Advanced	25(.46)	13(.24)	7(.13)	10(.18)	.22
	Exited	180(.70)	35(.14)	15(.06)	26(.10)	.08
	Native	403(.60)	84(.13)	89(.13)	94(.14)	.18

To test the degree of difference in the L/H conditional probabilities for grade 5, Table 12 reports the results from the critical ratio with chi square test of significance for each of the ELL groups and exited vs. native speakers. For both multiple choice and constructed response, the beginner and intermediate are significantly different than native English speakers while the advanced group was not different from native speakers for either set of results. Therefore, although the contrast of the regression βs is not seen as significantly different for the beginners and non-ELLs on the constructed response subtest, there continues to be greater apparent misclassification (based on P(L/H)s) of these data. Finally, of interest is the very large difference between exited and native speakers, with the probability of exited students being significantly lower than native speakers.

Table 12. Grade 5 Total Sample Conditional Probability P(L/H) Comparisons							
		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Beg vs Nat	0.464	0.151	3.078	9.473	1	0.002
	Int vs Nat	0.140	0.054	2.612	6.821	1	0.009
	Adv vs Nat	0.050	0.056	0.885	0.783	1	0.376
	Exit vs Nat	-0.019	0.021	-0.941	0.886	1	0.347
Constructed Response	Beg vs Nat	0.444	0.172	2.581	6.664	1	0.010
	Int vs Nat	0.282	0.070	4.027	16.219	1	0.000
	Adv vs Nat	0.038	0.075	0.504	0.254	1	0.614
	Exit vs Nat	-0.104	0.026	-4.031	16.248	1	0.000
N = 1175							

b. Comparison of Quadrant Distributions as Conditioned on Ancillary Variables

As noted above, several pieces of additional data were collected as part of the VAELL project. Included were data on some ancillary variables whose impact researchers worked hard to minimize in the project. Of the ancillary variables that were collected, reading, context, testwiseness, and a composite variable of five psychosocial questions were inspected in this study to see if they continue to be problematic for ELLs and others.

For the purposes of this study, only those students whose teachers indicated they had problems with the identified variable were included in the analyses. Once this subgroup of students was identified, the quadrant data were computed by item type and grade for each of the variables. Because of low sample sizes, an ELL group made up of the beginner, intermediate and advanced groups were used in the tests for significance, in addition to exited and native English speakers. The distribution data will be reported by each ELL group as well as by composite ELL group.

The student information for these variables was collected on the teacher questionnaire. For reading, teachers were asked to rate the student's ongoing reading proficiency in the classroom one a 1-5 scale, one being reading consistently below grade level to five, consistently above grade level. These analyses used levels 1 and 2 of this data, that is, reading consistently and sometimes below grade level as defined by the Maryland content standards and achievement levels. The context variable asked teachers if they believed students often had trouble accessing the context generally used in test items—be they textbook test items, standardized tests, or other types of tests the teachers use in their classroom. The dichotomous testwiseness variable asked about lack of familiarity with typical item and response formats used on tests, seeing many items on one page, bubbling, or using a separate answer sheet. If the teachers answered 'yes' to either the context or testwiseness question, those student data were included here. Finally, the psychosocial variable is a composite of five dichotomous questions: frustration, test anxiety, fatigue, distractibility, and lack of motivation. If students had problems with five, four or three of these variables, they were included in the following analyses.

The hypothesis is that the reading and the context variables will still continue to impact the ELL students with less language proficiency more than native speakers. It is unclear if the other variables will still differentially impact the less proficient ELL students as compared to other groups.

b.1 Reading

Grade 3 Table 13 presents the quadrant results, by group, for students whose teachers thought they were consistently or sometimes below grade level standards in their reading proficiency. Table 14 displays the critical ratio tests. For the multiple choice

results, conditional probability of L/H for beginner and intermediate students seems quite a bit higher than for native English speakers and chi-square findings indicate that there are significantly more ELL students in the L/H quadrant than native speakers (but not exited). The L/H conditional probabilities of the constructed response results for those students as well as advanced are higher than both control groups, and the chi-square tests report that ELLs have significantly more students with reading problems in the L/H quadrant than do either the exited or native speakers.

Table 13. Reading Distributions by Quadrant, Grade 3

	Group	HH	HL	LH	LL	P(L/H)
	Beginner	10(.20)	20(.41)	5(.10)	14(.29)	.33
Multiple	Intermediate	71(.41)	50(.29)	26(.15)	26(.15)	.27
Choice	Advanced	30(.68)	8(.18)	3(.07)	3(.07)	.09
	ELL	*135(.51)*	*62(.23)*	*39(.15)*	*30(.11)*	*.22*
	Exited	62(.57)	22(.20)	13(.12)	11(.10)	.17
	Native	179(.57)	84(.26)	25(.08)	31(.10)	.12
	Beginner	6(.12)	3(.06)	7(.14)	33(.67)	.54
Constructed	Intermediate	25(.15)	30(.17)	50(.29)	68(.39)	.67
Response	Advanced	16(.36)	5(.11)	13(.30)	10(.23)	.45
	ELL	*47(.18)*	*38(.14)*	*70(.26)*	*111(.42)*	*.60*
	Exited	57(.53)	13(.12)	20(.19)	19(.17)	.26
	Native	103(.32)	42(.13)	88(.28)	86(.27)	.46

Table 14. Reading Conditional Probability P(L/H) Comparisons, Grade 3

		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat vs ELL	-0.102	0.039	-2.600	6.761	1	0.009
	Exit vs ELL	-0.051	0.054	-0.942	0.887	1	0.346
Constructed Response	Nat vs ELL	-0.138	0.045	-3.035	9.212	1	0.002
	Exit vs ELL	-0.339	0.050	-6.775	45.900	1	0.000

Grade 5 Tables 15 and 16 report the fifth grade results of the quadrants by group who teachers said are consistently or sometimes below reading proficiency. As with grade 3, the composite ELL results will also be included. Conditional probability results

appear to be similar to what occurred for grade 3, and for multiple choice, ELL students with reading difficulties are significantly more apt to be in the L/H quadrant as compared to either exited or native speakers. Constructed response results for exited vs. native are similar as well. However, for the constructed response subtest, there does not appear to be a significant difference between the percentage of ELLs vs. native speakers in this quadrant.

Table 15. Reading Distributions by Quadrant, Grade 5

	Group	HH	HL	LH	LL	P(L/H)
	Beginner	1(.02)	24(.59)	5(.12)	11(.27)	.83
Multiple	Intermediate	25(.21)	56(.46)	12(.10)	29(.24)	.32
Choice	Advanced	15(.43)	14(.40)	3(.09)	3(.09)	.17
	ELL	*41(.21)*	*94(.48)*	*20(.10)*	*43(.22)*	*.33*
	Exited	61(.52)	41(.35)	8(.07)	7(.06)	.12
	Native	152(.48)	106(.33)	23(.07)	39(.12)	.13
	Beginner	1(.02)	6(.15)	3(.07)	31(.76)	.75
Constructed	Intermediate	17(.14)	46(.38)	12(.10)	47(.39)	.41
Response	Advanced	10(.29)	13(.37)	2(.06)	10(.29)	.17
	ELL	*28(.14)*	*65(.33)*	*17(.09)*	*88(.44)*	*.38*
	Exited	50(.42)	33(.28)	8(.07)	26(.22)	.14
	Native	105(.33)	80(.25)	43(.13)	92(.29)	.29

Table 16. Reading Conditional Probability P(L/H) Comparisons, Grade 5

		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat vs ELL	-0.196	0.065	-3.008	9.048	1	0.003
	Exit vs ELL	-0.212	0.071	-2.968	8.810	1	0.003
Constructed Response	Nat vs ELL	-0.087	0.081	-1.072	1.150	1	0.284
	Exit vs ELL	-0.240	0.085	-2.812	7.909	1	0.005

b.2 Context

Tables 17 and 18 display results for grade 3 while tables 19 and 20 illustrate findings from grade 5. For both grades the results are similar to those found in reading. That is, three of the four comparisons

are significant in each grade. Both multiple choice and constructed response results indicate a significant difference between ELLs and native speakers in grade 3 but only multiple choice is significant in grade 5. For exited vs. ELLs, probabilities are significantly different for only constructed response in grade 3 but both subtests in grade 5.

Grade 3

Table 17. Context Distributions by Quadrant, Grade 3

	Group	HH	HL	LH	LL	P(L/H)
	Beginner	9(.33)	9(.33)	5(.19)	4(.15)	.36
Multiple	Intermediate	50(.48)	25(.24)	18(.17)	12(.11)	.27
Choice	Advanced	17(.68)	4(.16)	4(.16)	0(.00)	.19
	ELL	*76(.49)*	*38(.24)*	*27(.17)*	*16(.10)*	*.26*
	Exited	45(.76)	4(.07)	9(.15)	1(.02)	.17
	Native	113(.66)	26(.15)	21(.12)	11(.06)	.16
	Beginner	4(.15)	2(.07)	4(.15)	17(.63)	.50
Constructed	Intermediate	11(.11)	22(.21)	30(.29)	42(.40)	.73
Response	Advanced	7(.29)	4(.37)	8(.06)	6(.29)	.53
	ELL	*22(.14)*	*28(.18)*	*42(.27)*	*65(.41)*	*.66*
	Exited	31(.53)	7(.12)	10(.17)	11(.19)	.24
	Native	68(.40)	19(.11)	37(.22)	47(.28)	.35

Table 18. Context Conditional Probability P(L/H) Comparisons, Grade 3

		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat vs ELL	-0.105	0.054	-1.970	3.880	1	0.049
	Exit vs ELL	-0.095	0.067	-1.431	2.048	1	0.152
Constructed Response	Nat vs ELL	-0.304	0.075	-4.025	16.204	1	0.000
	Exit vs ELL	-0.412	0.090	-4.604	21.194	1	0.000

Grade 5

Table 19. Context Distributions by Quadrant, Grade 5

	Group	HH	HL	LH	LL	P(L/H)
	Beginner	0(.00)	11(.65)	0(.00)	6(.35)	0 High ratings
Multiple	Intermediate	7(.13)	31(.55)	7(.13)	11(.20)	.50
Choice	Advanced	2(.17)	6(.50)	2(.17)	2(.17)	.50
	ELL	*9(.11)*	*48(.57)*	*9(.11)*	*19(.22)*	*.50*
	Exited	31(.60)	16(.31)	2(.04)	3(.06)	.06
	Native	73(.48)	46(.31)	11(.07)	20(.13)	.13
	Beginner	0(.00)	3(.18)	0(.00)	14(.82)	0 High ratings
Constructed	Intermediate	6(.11)	22(.39)	5(.09)	23(.41)	.46
Response	Advanced	2(.17)	4(.33)	0(.00)	6(.50)	.00
	ELL	*8(.09)*	*29(.34)*	*5(.06)*	*43(.51)*	*.39*
	Exited	25(.48)	11(.21)	2(.04)	14(.27)	.07
	Native	57(.38)	29(.19)	17(.11)	47(.31)	.23

Table 20. Context Conditional Probability P(L/H) Comparisons, Grade 5

		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat vs ELL	-0.369	0.123	-2.989	8.935	1	0.003
	Exit vs ELL	-0.439	0.125	-3.516	12.365	1	0.000
Constructed Response	Nat vs ELL	-0.155	0.144	-1.079	1.165	1	0.281
	Exit vs ELL	-0.311	0.144	-2.156	4.648	1	0.031

b.3 Testwiseness

Tables 21 and 22 display results for grade 3 while tables 23 and 24 illustrate findings from grade 5. The probability distributions in both grades indicate that misclassified n's are generally quite small, and the only significant differences in the comparison of the L/H quadrants between ELL and the both control groups occurs for the constructed response subtest in grade 3.

Grade 3

Table 21. Testwiseness Distributions by Quadrant, Grade 3

	Group	HH	HL	LH	LL	P(L/H)
	Beginner	2(.18)	4(.36)	2(.18)	3(.27)	.50
Multiple	Intermediate	28(.51)	13(.24)	7(.13)	7(.13)	.20
Choice	Advanced	7(.64)	2(.18)	2(.18)	0(.00)	.22
	ELL	*37(.49)*	*18(.24)*	*11(.15)*	*10(.13)*	*.23*
	Exited	28(.74)	4(.11)	6(.16)	0(.00)	.18
	Native	62(.64)	19(.20)	10(.11)	6(.06)	.14
	Beginner	0(.00)	0(.00)	3(.27)	8(.73)	1.00
Constructed	Intermediate	7(.13)	14(.26)	14(.26)	20(.36)	.67
Response	Advanced	2(.18)	2(.18)	5(.46)	2(.18)	.71
	ELL	*9(.12)*	*16(.21)*	*22(.29)*	*29(.38)*	*.71*
	Exited	20(.53)	6(.16)	6(.16)	6(.16)	.23
	Native	38(.39)	7(.07)	24(.25)	28(.29)	.39

Table 22. Testwiseness Conditional Probability P(L/H) Comparisons, Grade 3

		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat vs ELL	-0.090	0.073	-1.235	1.526	1	0.217
	Exit vs ELL	-0.053	0.089	-0.591	0.349	1	0.555
Constructed Response	Nat vs ELL	-0.323	0.102	-3.152	9.936	1	0.002
	Exit vs ELL	-0.479	0.116	-4.126	17.022	1	0.000

Grade 5

Table 23. Testwiseness Distributions by Quadrant, Grade 5

	Group	HH	HL	LH	LL	P(L/H)
	Beginner	0(.10)	6(.60)	0(.00)	4(.40)	0 High ratings
Multiple	Intermediate	4(.15)	17(.63)	3(.11)	3(.11)	.43
Choice	Advanced	1(.13)	4(.50)	1(.13)	2(.25)	.50
	ELL	*5(.11)*	*27(.60)*	*4(.09)*	*9(.20)*	*.44*
	Exited	15(.58)	8(.31)	2(.08)	1(.04)	.12
	Native	33(.40)	34(.41)	9(.11)	7(.09)	.21
	Beginner	0(.00)	1(.10)	0(.00)	9(.90)	0 High ratings
Constructed	Intermediate	3(.11)	14(.52)	1(.04)	9(.33)	.25
Response	Advanced	1(.13)	3(.38)	0(.00)	4(.50)	.00
	ELL	*4(.09)*	*18(.40)*	*1(.02)*	*22(.49)*	*.20*
	Exited	12(.46)	7(.27)	1(.04)	6(.23)	.08
	Native	30(.36)	16(.19)	9(.11)	28(.34)	.23

Table 24. Testwiseness Conditional Probability P(L/H) Comparisons, Grade 5							
		diff	se diff	critical ratio	Chi-square	df	P
Multiple Choice	Nat vs ELL	-0.230	0.177	-1.298	1.685	1	0.194
	Exit vs ELL	-0.327	0.183	-1.784	3.184	1	0.074
Constructed Response	Nat vs ELL	0.031	0.191	0.161	0.026	1	0.872
	Exit vs ELL	-0.123	0.194	-0.636	0.404	1	0.525

b.4 Psychosocial

The findings for the psychosocial variable are similar to testwiseness for both grades (see Tables 25 and 26 for grade 3 and tables 27 and 28 for grade 5). That is, the probability distributions in both grades indicate that misclassified n's are generally quite small, and the only significant differences in the comparison of the L/H quadrants between ELL and the both control groups occurs for the constructed response subtest in grade 3.

Grade 3

Table 25. Psychosocial Distributions by Quadrant, Grade 3

	Group	HH	HL	LH	LL	P(L/H)
	Beginner	8(.50)	4(.25)	1(.06)	3(.19)	.11
Multiple	Intermediate	33(.56)	8(.14)	14(.24)	4(.07)	.30
Choice	Advanced	12(.67)	3(.17)	3(.17)	0(.00)	.20
	ELL	*53(.57)*	*15(.16)*	*18(.19)*	*7(.08)*	*.25*
	Exited	30(.73)	5(.12)	4(.10)	2(.05)	.12
	Native	75(.64)	19(.16)	16(.14)	7(.06)	.18
	Beginner	2(.13)	1(.06)	3(.19)	10(.63)	.60
Constructed	Intermediate	11(.17)	11(.17)	20(.34)	17(.29)	.65
Response	Advanced	5(.28)	4(.22)	4(.22)	5(.28)	.44
	ELL	*18(.19)*	*16(.17)*	*27(.29)*	*32(.34)*	*.60*
	Exited	20(.49)	6(.15)	6(.15)	9(.22)	.23
	Native	42(.36)	15(.13)	27(.23)	33(.28)	.39

Table 26. Psychosocial Conditional Probability P(L/H) Comparisons, Grade 3							
		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat vs ELL	-0.078	0.065	-1.191	1.418	1	0.234
	Exit vs ELL	-0.136	0.076	-1.797	3.228	1	0.072
Constructed Response	Nat vs ELL	-0.209	0.094	-2.227	4.958	1	0.026
	Exit vs ELL	-0.369	0.110	-3.348	11.211	1	0.001

Grade 5

Table 27. Psychosocial Distributions by Quadrant, Grade 5

	Group	HH	HL	LH	LL	P(L/H)
	Beginner	0(.00)	9(.64)	0(.00)	5(.36)	0 High ratings
Multiple	Intermediate	12(.25)	24(.50)	3(.06)	9(.19)	.20
Choice	Advanced	4(.33)	4(.33)	2(.07)	2(.17)	.33
	ELL	*16(.22)*	*37(.50)*	*5(.07)*	*16(.22)*	*.24*
	Exited	26(.53)	18(.37)	2(.04)	3(.06)	.07
	Native	82(.54)	42(.28)	10(.07)	19(.12)	.11
	Beginner	0(.00)	3(.22)	0(.00)	11(.77)	0 High ratings
Constructed	Intermediate	9(.19)	18(.38)	4(.08)	17(.35)	.31
Response	Advanced	3(.25)	3(.25)	0(.00)	6(.50)	.00
	ELL	*12(.16)*	*24(.32)*	*4(.05)*	*34(.46)*	*.25*
	Exited	19(.39)	14(.29)	4(.08)	12(.25)	.17
	Native	63(.41)	26(.17)	19(.12)	45(.29)	.23

Table 28. Psychosocial Conditional Probability P(L/H) Comparisons, Grade 5							
		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat vs ELL	-0.129	0.098	-1.314	1.728	1	0.189
	Exit vs ELL	-0.167	0.105	-1.589	2.524	1	0.112
Constructed Response	Nat vs ELL	-0.018	0.118	-0.155	0.024	1	0.877
	Exit vs ELL	-0.076	0.134	-0.568	0.322	1	0.570

c. Summary of Quadrant Distribution Results

Total sample results

It appears that the amount of misclassification becomes less as ELL students attain more proficiency in English. The advanced ELL level seems to be at a turning point for determining more vs. relatively fewer inconsistencies in the L/H quadrants, and the point where substantially fewer differences between ELL groups either the exited ELL students or native speakers occur.

In both grades, beginner and intermediate ELLs are always significantly more represented in L/H quadrant than native English speakers, for both multiple choice and constructed response. For beginners, results indicate that, for the multiple choice subtest, 29% and 55% of students in 3rd and 5th grades, respectively, scored very low on the benchmark test while teachers reported that these students had some ability in mathematics. For constructed response, the percentages (53% and 63%, respectively) were even higher. This is compared to an 8% inconsistent classification for native speakers on the multiple choice subtest (both grades), and 28% (in grade 3) and 18% (in grade 5) for this group on the constructed response subtest. On the other hand, the degrees of inconsistency in the L/H quadrant for advanced ELL students are significantly different from native speakers only for the third grade constructed response subtest.

Of interest, exited ELL students tend to have less L/H classification inconsistencies than native speakers for constructed response subtest in both grades. That is, they seem to flourish with this approach while it appears that native students have more problems.

Results by ancillary variables

Reading, context and psychosocial all have significant differences between ELL and native English speakers for 3rd and 5th grade students on the multiple choice subtests, and for constructed response in 3rd grade. Also, there were significant differences between the amount of exited students in L/H in every case but 3rd multiple choice as compared to ELLs (who were always lower). Testwiseness was the one variable that, on the whole, seemed to be leveled in its effect on scores across groups for this quadrant. The only exception to this is in the 3rd grade constructed response subtest, where ELL misclassification percentages are significantly higher than the percentages for native English speakers and exited.

It is important to keep in mind that, for context, psychosocial and testwiseness, the n's for the inconsistencies in this quadrant are not large, though they are consistently larger for third grade than fifth. This suggests that these variables might tend to interact with reading and/or developmental issues.

Discussion

So, what does this mean? Several implications seem to arise from these data. Of these, six will be discussed below.

First, it is important to consider that the amount of non-parity still evidenced could occur for a few different reasons. Among them may be:

- Teachers may be worse at estimating the mathematics abilities of ELL students (as defined in the teacher questionnaire items) than they are at estimating these abilities for native English speakers and exited students.

- Teachers may be generally correct but one or more types of OTL may be a problem in classroom of ELLs—specifically, in addition to possible instructional disparities in ELL and mainstream classes around the scope or amount of time spend on mathematics teaching, the inconsistency may be due to teachers using more computational type problems and verbal discussion vs. word problems in the classroom.
- The assessment field still has work to do to make tests more accessible, especially for beginner and intermediate English learners.

We suspect that, while teachers may have some more reservations about the actual abilities of low literacy students, that it is not primarily the first bullet because of the consistent variation of ability across groups. Teachers generally identified low proficient ELs with less mathematics ability than their peers (which we think is reasonable), although they also identified some of the students with low literacy as having mathematics knowledge. That, is, they seemed to be able to discriminate within each of the ELL subgroups and this discrimination seemed to be similar to how they viewed the range of abilities of exited and native English speaker students.

On the other hand, we think there is a good chance that both of the other bullets are interacting with test score performance and these disparities seem to be particularly problematic for beginners and intermediate students. In the second bullet, the unequivalent instructional time may be present or equivalent time may be more often used to catch up from previous schooling deficiencies. Experience suggests additionally, however, that many teachers of ELLs will resort to disparities in instructional and assessment approaches (using computational problems more often than word problems)

as a way to communicate with the students who have less literacy and might be faced with word problem difficulties if the problems are not phrased, presented or scored correctly. The results suggest that, in fact, especially the beginner students actually prefer a chance to explain their thinking, assuming they understand the item requirements. Finally, the third bullet considers that, as professional assessment developers and consumers, we need to improve how we interpret the scores of the lower literacy ELL students while we work to upgrade how we might collect valid data from these students.

Second, as introduced above, the regression results suggest that teachers' ratings of mathematics knowledge for beginner students correspond poorly to test scores when measured with multiple choice items but show promise when measured with constructed response item formats (teachers didn't know which format we would use for particular skills and knowledge elements when they rated students on their levels of mathematics skills/elements). However, the constructed response quadrant results suggest there is still a large difference in misclassification for this group and item type as compared to native speakers and exited. Further, intermediate students struggle with both multiple choice and constructed response formats.

The multiple choice problem for beginner and intermediate students seem to signal the language issues which exist in multiple choice items but are frequently overlooked by mainstream test makers. This includes the language of discrimination which is necessary for choosing distractors. In addition to constructed response, there may be other item types or testing approaches that are an alternative to multiple choice for beginners and item type alternatives to both item types for intermediate ELL students. These item type alternatives should address some of the reading and writing language issues associated

with presenting item requirements to students, and increase the response options for students who must rely on strengths other than language. They may be items that use other close ended formats and/or are more interactive without resorting to language, such as those which could be computer based.

Third, the total sample quadrant analyses don't seem to get at the performance problems of Advanced ELLs for 3rd on the third grade multiple choice subtest (and to some extent at fifth grade). The n's for the ancillary variable analyses didn't support looking at advanced results separately. With the regression results closer to native speakers/exited for grade 5, as compared to grade 3 this may be a schooling and/or reading familiarity issue. That is, advanced ELLs at grade 5 may have had time to learn the academic language required to utilize the multiple choice items. At the same time, while there is some evidence in the regressions that their level of parity with the control groups is somewhat higher for constructed response subtests than multiple choice at this grade, they, like their exited peers, (and beginners, to some extent) may prefer constructed response. This group may or may not prefer the constructed response format in grade 3, but they still seem to have trouble with it.

Fourth, ELL experts suggest that, as the students learn English, they go through a phase where they become hyper-sensitive to inevitable textual inconsistencies (which can occur for several reasons). This is compared to their beginner peers who are focusing on main themes in items and don't see the inconsistencies, or to their exited peers and native speakers who know how to gloss over these inconsequential textual elements. This may partially explain why intermediate students at both grades, and advanced at grade 3 do not seem to respond as well as exited/native. The evidence may also point to why

advanced ELLs seem to be a transition group—as the students' language matures, they overcome this hyper-vigilant hump and become more similar to their peers who are more language proficient.

Fifth, the constructed response quadrant analyses report that, consistently, there are significantly more correctly classified Exited students than native speakers. These results occur when the total sample differences were investigated as well as when ancillary constructed response results were isolated for students with access needs. Further, the R square coefficients in the regression results are more similar to natives for the constructed response subtest results in both grades (with higher coefficients in grade 5), as compared the results for multiple choice scores (although the betas are not significantly different). This may suggest that exited students also benefit from the open-ended quality of this item type which allows them to explain what they know. Thus, language may still be an issue for these students, even though they "do all right" with the multiple choice format.

Finally, regarding the ancillary results: consistently, reading load is still an issue in these items for ELLs. Context also appears to affect younger students—this may be a result of less experience in US schools, and with US culture and experiences. Context for older students, as well as psychosocial variables continue to affect the inconsistent classification results significantly more so for ELLs than with native or exited students. While the n's aren't large as they are for reading or for context in the third grade, it should be remembered the control contrasts were only for native and exited students whose teachers said they had problems. It is likely the contrasts would be starker between ELLs, exited and natives with access needs, versus other native English speakers or exited students. As such, in order to reach parity with the validity inferences in these

three ancillary variables, the findings suggest that there is still work to do in minimizing the ancillary impact of the variables for ELLs and native English speakers with problems accessing test items. Therefore, while the n's in two of the variables (as well as testwiseness) are rather small individually, the findings suggest that, taken as a whole, there is a reasonably large group of students in most large-scale test taking situations who can benefit from additional improvements. Ancillary variables such as these suggest a few points of focus for doing this work. Some of these elements may aid various students with language difficulties, while some will undoubtedly be associated with unique ELL variables which affect language acquisition and cultural experiences.

References

Kopriva, R.J., & Mislevy, R. (2005). *Final research report of the Valid Assessment of English Language Learners Project* (C-SAVE Rep. No. 259). USED, IES and Madison, WI: University of Wisconsin, Center for the Study of Assessment Validity and Evaluation.

Koran, J. & Kopriva, R.J. (2007 in press), Proper assignment of accommodations to individual students. In Kopriva, R.J., *Improving Testing for English Language Learners: A Comprehensive Approach to Designing, Building, Implementing, and Interpreting Better Academic Assessments*, Erlbaum/Routledge Publishers, NY, NY.

Grade 3 MC

Figure 1. Beginners

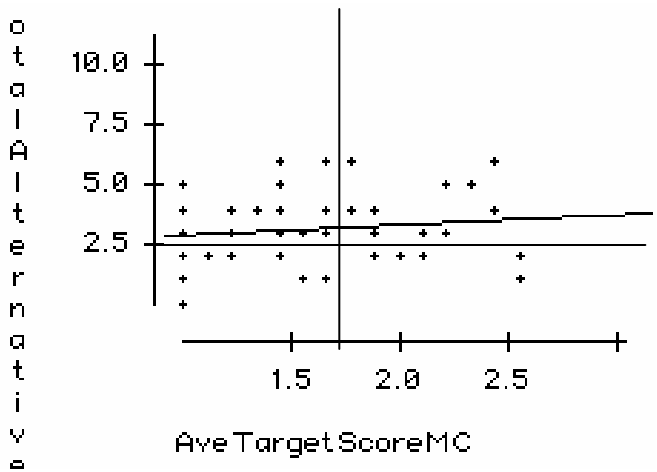


Figure 2. Intermediate

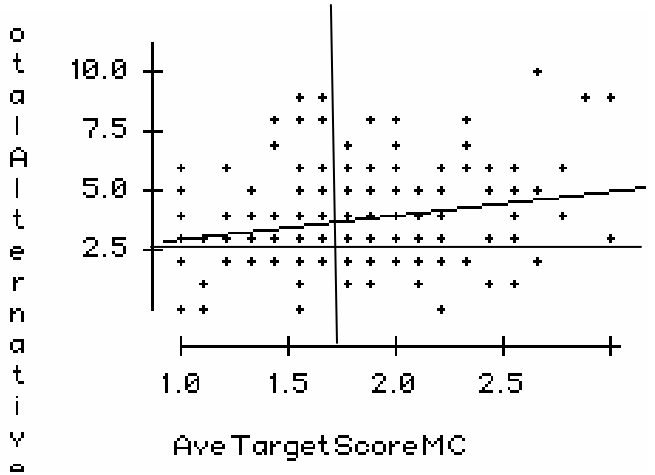


Figure 3. Advanced

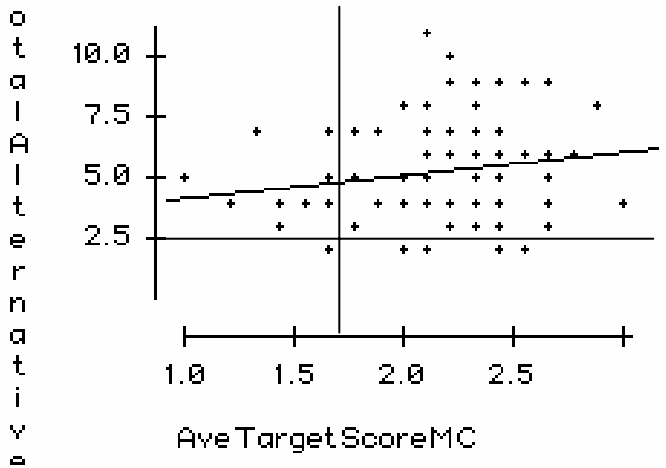


Figure 4. Exited

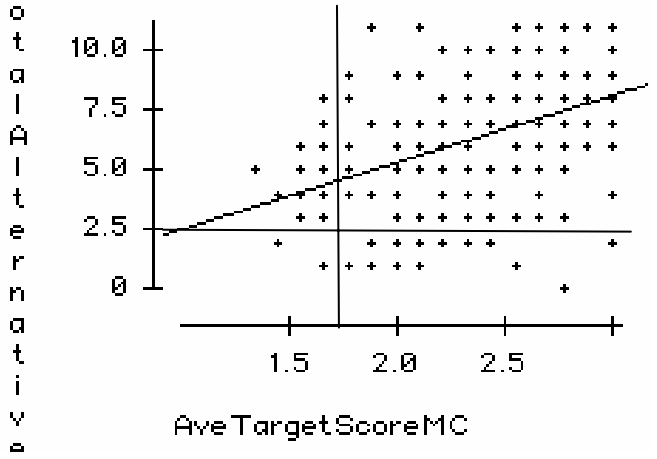
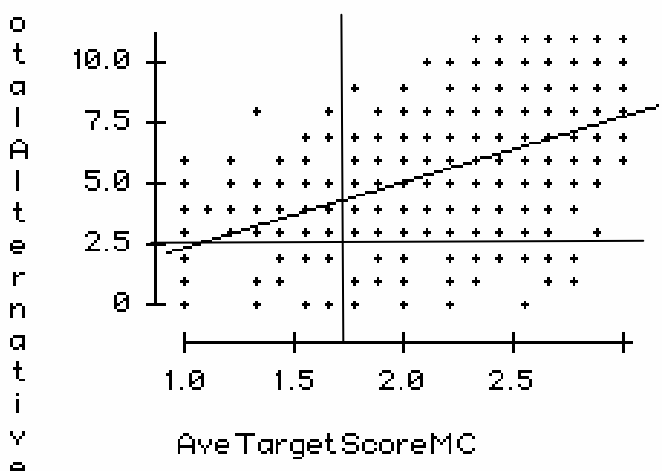


Figure 5. Native Speakers



Grade 3 CR

Figure 6. Beginners

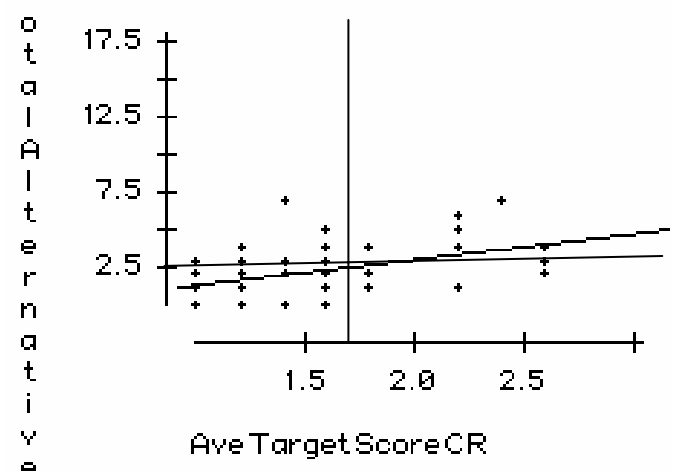


Figure 7. Intermediate

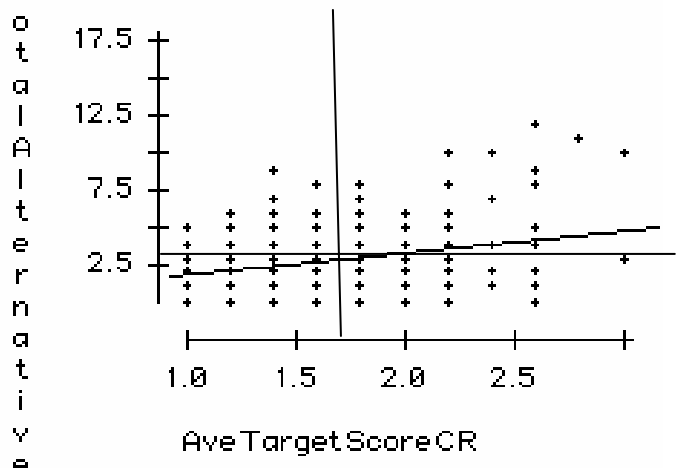


Figure 8. Advanced

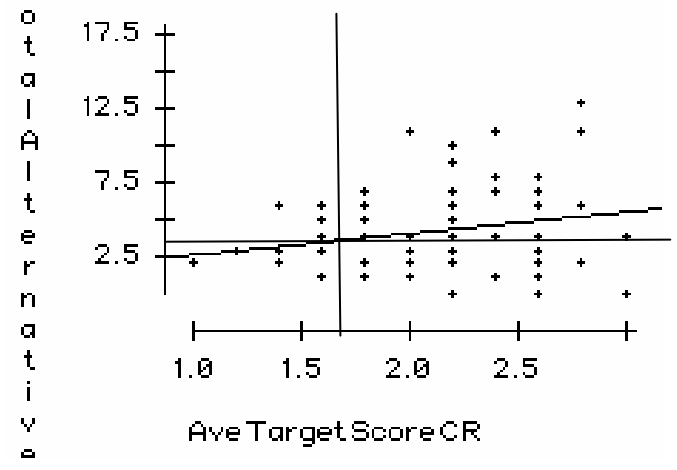


Figure 9. Exited

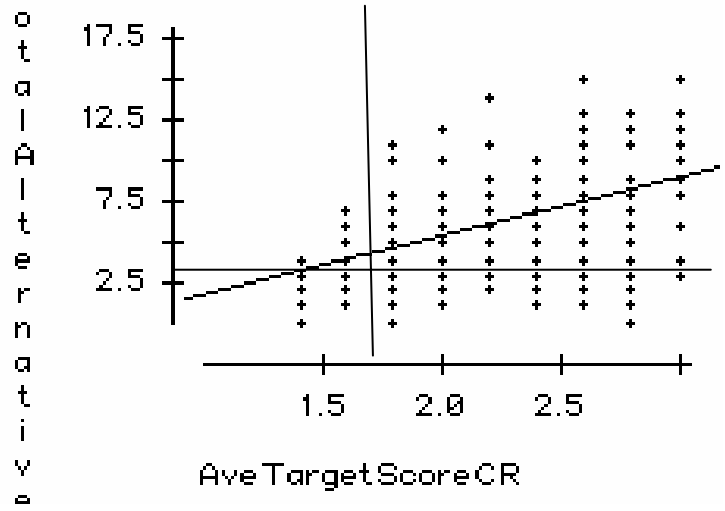
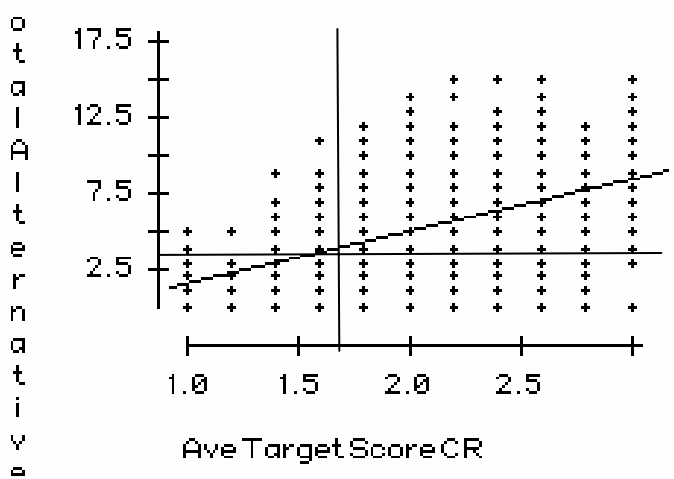


Figure 10. Native Speakers



Grade 5 MC

Figure 11. Beginners

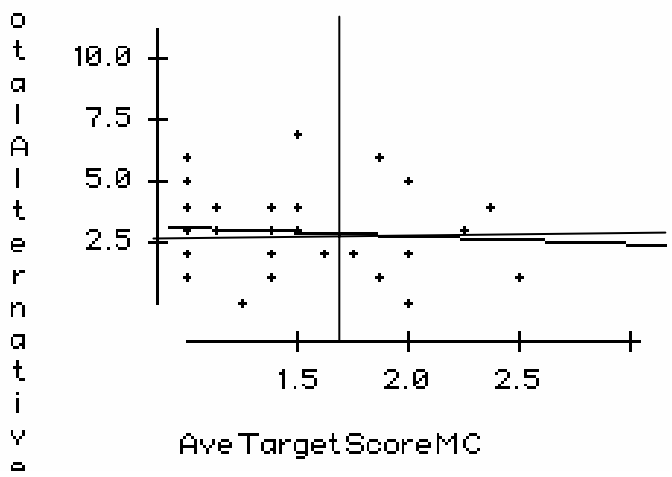


Figure 12. Intermediate

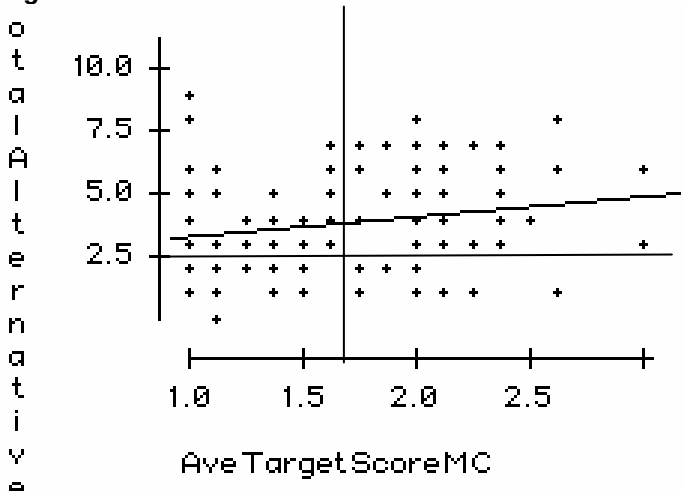


Figure 13. Advanced

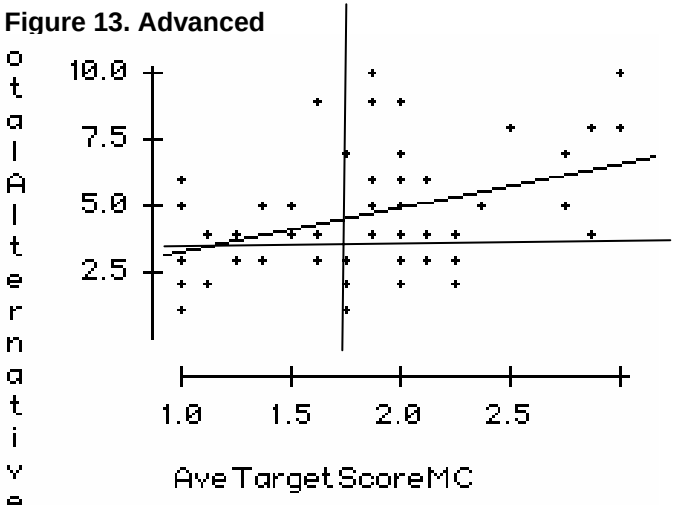


Figure 14. Exited

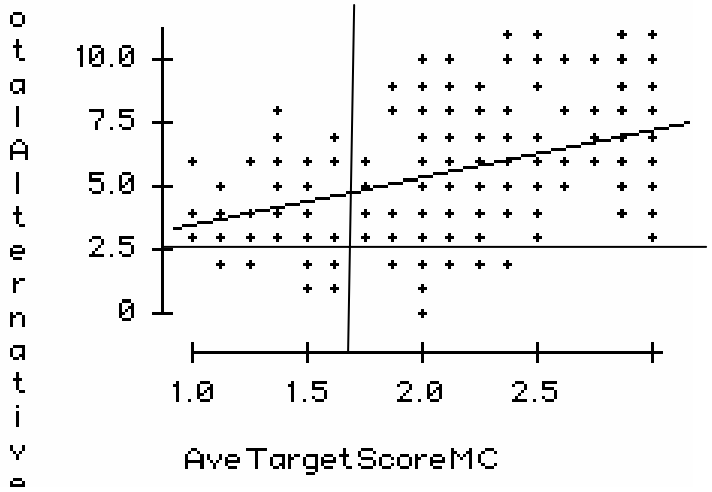
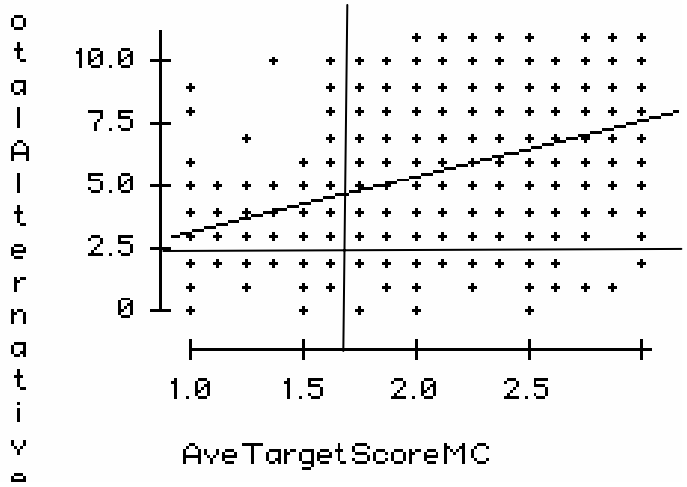


Figure 15. Native Speakers



Grade 5 CR

Figure 16. Beginners

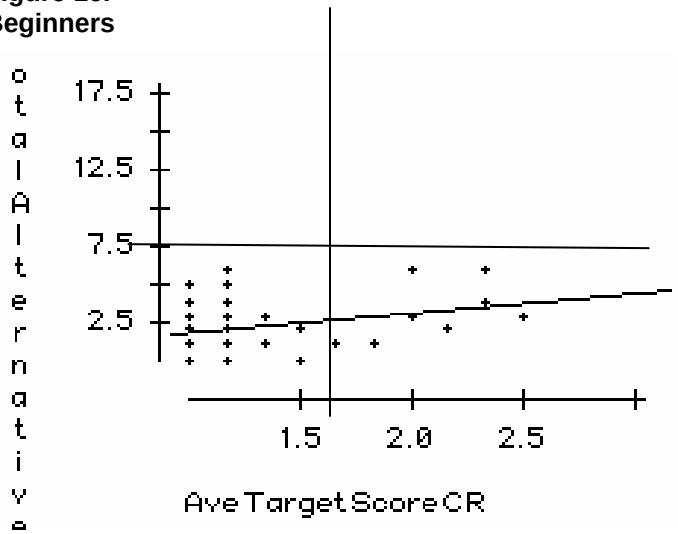


Figure 17. Intermediate

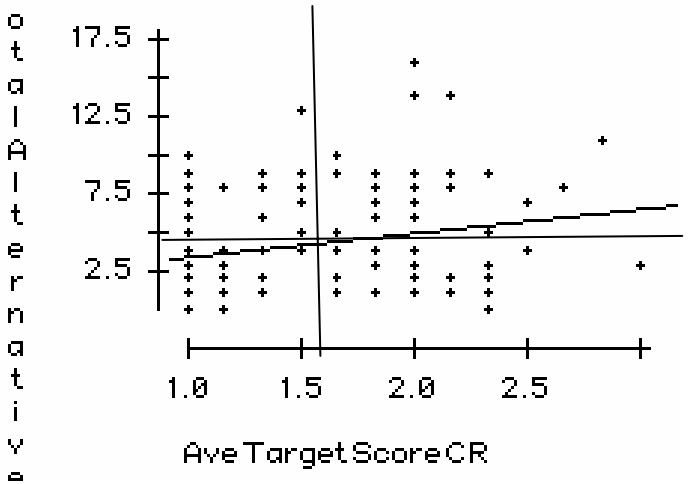


Figure 18. Advanced

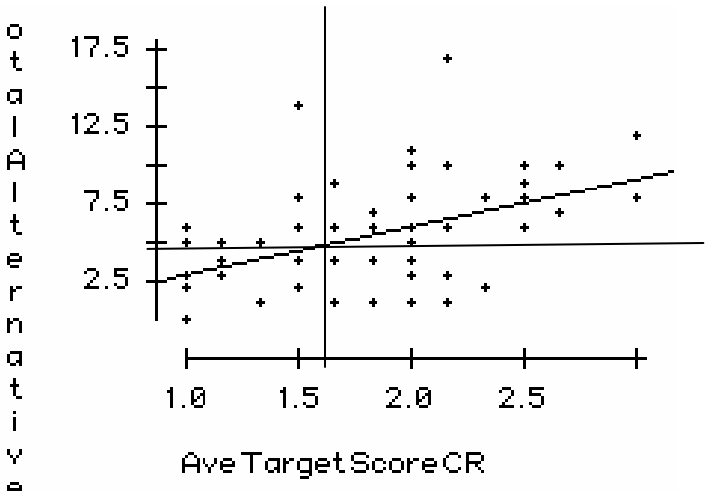


Figure 19. Exited

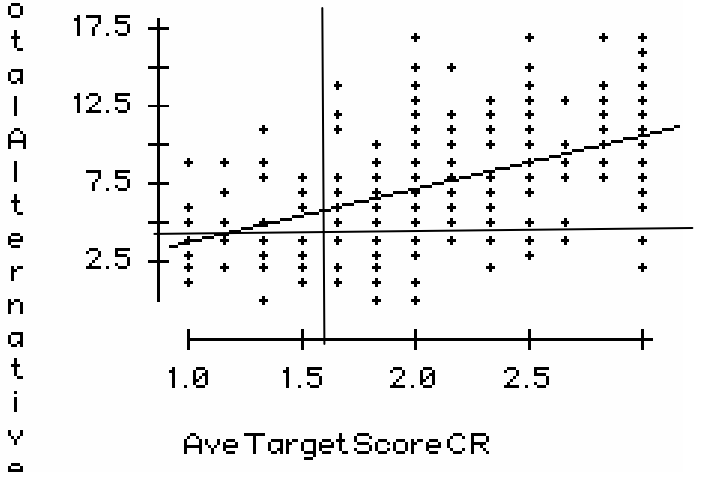


Figure 20. Native Speakers