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 by evaluating the effect this has on the validity of the score inferences for control groups and ELL students with different needs, strengths, and levels of language proficiency. Given the individualized application of accommodations to ELL students who needed them, it was expected that the validity of the score inferences would be similar for 3rd and 5th grade English language learners and the control groups of exited and native English speakers. In all, data from 19 multiple choice and constructed response mathematics items per grade were collected and regressed on student data from a criterion teacher rating measure in mathematics. Results indicated relationships between the multiple choice and criterion measure findings for ELLs were generally very poor compared to the control groups, but the comparisons for the constructed response data were more promising, especially for beginning ELL students in both grades and advanced ELL 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 identified as knowing at least some mathematics on the criterion measure. This study raises many questions about the validity of inferences drawn from large-scale assessments when the academic achievement of students with lower English proficiency is measured using traditional multiple choice formats in particular (even with promising accommodations). The data underscore the importance of response accommodations for at least some of these students, and suggest that item type may interact with grade level of the test takers.

Recent shifts towards inclusion of English language learners (ELLs) in large-scale testing programs have resulted in concerns about the validity of inferences being drawn from test scores. These concerns are related to the potentially confounding relationship between proficiency in the English language and literacy, and ELL's ability to properly be able to demonstrate their academic knowledge and skills. Additionally there is the recognized cultural disconnect frequently experienced by English learners (e.g. Solano-Flores & Trumbull, 2003; Kopriva, 2000; 2007). With current federal legislation and the 71.9% increase of ELLs in K-12 U.S. public schools over the last decade (cite), this means that virtually all states are impacted, and that they are responsible for properly testing all English learners at identified grades.

Over the last 10 years, the primary way that educators and researchers have attempted to address the validity of large-scale assessments for both ELLs and students with disabilities is by providing appropriate accommodations. While much more research is still required to confirm the efficacy of specific accommodations, some progress has been made. Over time, the most relevant accommodations for ELLs have been identified, and some work has been completed to determine which accommodations might be effective, when, and for whom.

All in all, approximately 12 overlapping presentation, administration, and response accommodations seem to be useful, especially for ELL students with some proficiency in English. The accommodations include increasing accessibility of the assessment text and accommodations which provide some language supports (Pennock-Roman & Rivera; 2006). Work on better editing techniques and increasing methods of

access in the items (usually through visual displays), if handled correctly, seem to be promising for ELLs while not being problematic for native English speakers (e.g Abedi, 2001; Kopriva & Cameron, 2007; Carr & Kopriva, 2007). Language support materials, for instance bilingual glossaries and L1 language aid, have been more deliberately specified (e.g. Abedi, Courtney, Mirocha, Leon & Goldberg, 2001; Robin, Sireci & Hambleton, 2003; Duncan, del Rio Parent, Chen, Ferrara, Johnson, Oppler, & Shieh, 2005), and a small number of administration accommodations have been identified as being promising in supporting the language and cultural improvements. All in all, research on accommodations is rather well known and can be accessed through a number of literature reviews (for example see Sirici, Li, & Scarpati, 2003; Pennock,-Roman & Rivera, 2006; Kopriva, 2007).

On a related front, it has been widely thought that effectiveness is conditioned by matching student needs and proper accommodations, but, until recently, little work was completed to assess whether this was being done and the extent to which it mattered. Improvements in matching for students with disabilities have been ongoing the last several years (for instance see Fuchs, Fuchs, Eaton, Hamlett, & Karns ,2000; Fuchs & Fuchs, 2006; Tindal, 2006). While they didn't give specific guidance, Butler and Stevens (1997) identified three student factors that appeared to be salient for differentiating accommodations for English learners: English language proficiency, prior formal schooling, and length of time in U.S. Further, Rivera, Collum, Shafer, Willner, and Sia (2006) recently identified direct and indirect language signposts that could help teachers better link student needs and test options.

Two related studies with ELLs suggest that matching does have an effect and that some methods appear to be more salient than others. In the first study (Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007), researchers randomly assigned 3rd and 4th grade ELLs (n=271) to one of 3 conditions [the "correct" package of accommodations based on the assignments by a version of the STELLA matching system, no accommodations, or an incorrect or incomplete package of accommodations]. Accommodations selected for this study were the bilingual word list, picture to word list, and oral English, and the packages consisted of one, two or three accommodations. The STELLA system (Selection Taxonomy for English Language Learner Accommodations) is a recently completed computerized system designed to match ELL student needs and accommodations, and this study focused on select decision branches of the complete system (see Kopriva, Chapter 10, 2007). Results indicated that student performed significantly better on test items when the package of accommodations was individualized and when their unique needs as ELLs were properly considered. Of interest, students who received the incorrect or incomplete package did not score differently than those who received no accommodations.

In the second study, Koran & Kopriva (2006) researched how the STELLA findings compared to a random assignment of accommodations for each student and three sets of teacher recommended assignments. The three sets of teacher recommendations were compiled before teachers utilized the STELLA approach, after completing student data forms for STELLA, and an additional recommendation based on elements associated with factors not considered in STELLA, but important to the teacher. After reviewing student data from STELLA forms for 114 students in three states, four ELL practitioners blindly ranked the accommodation recommendations associated with each student,

ranking the five sets from best to worse fit as compared to data retrieved from the forms. Goodness of fit results indicated that STELLA was found to consistently and significantly be the best fit, over students, and over all recommendation alternatives. On the other hand, none of teacher recommendations proved to be different from the random assignment of accommodations or from each other. This is significant because teacher recommendations are currently the method used most often to assign ELL students to accommodations.

These findings lend support for the importance of appropriately assigning accommodations and its central role in the on-going debate about accommodation use. Without more guidance, the current approach may be limiting the validity of the inferences for students even when accommodations are well designed.

This current study represents a status check on the effectiveness of efforts to improve the validity of test score inferences for ELLs to-date. As researchers and practitioners have improved the way ELL students are served in large-scale tests, both in terms of accommodations and assignment, little information has been compiled to evaluate if the same inferences can now be legitimately made from large-scale test scores for this population as compared to their native English speaking peers. In other words, is the relationship between student ability and test scores the same for identified groups of ELLs and non-ELLs? Over student needs and accommodations, is this population being properly served when quality language and administration accommodations are designed, implemented, and monitored, and when assignment follows the logic of that later quantified in the STELLA system? While not all accommodations were used here, seven

were chosen because they parsimoniously incorporated most of the elements from the accommodations identified as promising, and were considered to be "doable" for a wide range of states.

Specifically, three sets of research questions were asked:

1) Are the relationships between scores on large-scale mathematics 'testlets' and criterion mathematics measures designed to provide alternate estimates of student ability the same for ELLs, exited ELLs, and native English speakers? Are the relationships the same at grade 3 and grade 5, and for multiple choice and constructed response items?

2) Are the relationships the same for ELLs at each of the three levels of language proficiency (beginner, intermediate and advanced)?

3) If relationships are not the same, what is the extent of the problem for ELLs who are identified on the criterion as having some mathematics knowledge and skills, and how do these findings compare to those from native speakers and exited students who are identified as having at least some mathematics knowledge and skills?

The objective under question 3 is to determine the extent to which the number of students with little knowledge of mathematics could be influencing the results in questions 1 and 2, as compared to students who have mathematics knowledge and skills. As the accommodation improvements are designed to impact the students with knowledge in a content area, it was hypothesized that looking only at the relationship

of ELL and control students with mathematics skills may help to clarify the extent of the problem if there is one.

Method

This study is part of a larger project, The Valid Assessment of English Language Learners (VAELL, Kopriva & Mislevy, 2005). The data for the investigation were collected in the fall and winter of 2004/2005 in a large-scale testing setting. Empirically grounded 'plain language' or universal design mathematics test items were developed to provide enhanced access to students with language difficulties, two types of glossary lists, L1 support, and item-specific manipulatives designed to increase access through manipulation were created to be used with selected ELL students, and a small number 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 benchmark assessment, and implementation of supplementary test materials and administration accommodations was staffed by project personnel and monitored for quality control.

1. Sample

Usable data were obtained from 2420 third and fifth grade students from 21 schools in a school district in Maryland. Nineteen of the schools were selected as those who had high levels of English learners, and additional two schools were identified where student scores on previous large-scale tests were the highest in the district. 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, and other

language development variables (e.g., length of time in country, level of reading, writing and mathematics achievement, and demographic characteristics). Native English speakers and exited ELLs acted as control groups.

Table 1 lists the percent of students in the three ELL focal groups, and also exited and native English speaking students. Students with IEPs were not included in the data set, and neither were students whose parents chose to keep them out of ELL services (36 in grade 3 and 49 in grade 5). ELL students were identified by the district as beginner, intermediate or advanced based on an older English language proficiency test. As per the standards on this test, beginners were students with little vocabulary or oral skills, intermediate students were developing social English language and very basic literacy skills, and advanced students were socially conversant in English, and could read simple text. In the vocabulary of the new group of English language proficiency tests, these students would probably be called prefunctional or emerging, beginner, and intermediate, respectively (Abedi, 2007).

Table 1: Frequency (%) of Students in Each ELL Level

	Grade 3	Grade 5
Beginning	52 (4.2%)	46 (3.9%)
Intermediate	198 (15.9%)	148 (12.6%)
Advanced	75 (6.0%)	55 (4.7%)
Exited	245 (19.7%)	256 (21.8%)
Native English	675 (54.2%)	670 (57.0%)
	N = 1245	N = 1175

2. Instruments and Other Materials

Two instruments, three kinds of accommodation materials, and a set of project supports were developed for this project. First, eleven multiple choice items and eight constructed-response items in each grade were developed to be equivalent access-based versions of released mathematics items. These items were keyed to the Maryland's standards and indicators and tied to the district curriculum plan as coursework taught in the district's third and fifth grade classes by the time of testing. Once the items were completed and reviewed they were inserted into the district's large-scale benchmark test in mathematics which was designed to mirror the state's large-scale assessment. The administration also mimicked the state's testing procedures in an effort to provide students a school-wide practice trial prior to the official state testing window.

The access-based items were written in English and followed the template construction and access-based item writing guidelines later codified in Kopriva, 2007. Guidelines included guidance about how to retain the content and cognitive complexity of the existing items while minimizing language load and providing compensatory avenues where students could access the item requirements. Textual changes included the increased use of visuals in the prompts, stem, or response options, format options (for instance, use of bullets and attention to space), use of basic sentence and phrasal structures, and use of plain language words and phrases for non-targeted content. Accommodation tools, explained below, were designed to support the text. For the constructed response items, besides responding in the traditional written format (in English), students were also allowed to demonstrate their knowledge through use of illustrations, and native language and/or code-switching, using their L1 phonetic and literacy skills along with their emerging English capacity. During item development

mathematics experts reviewed the items to ensure that they were measuring the same targeted mathematics knowledge and skills as the original items, and district staff reviewed the items to ensure they were measuring the skills they were targeting. Items were piloted in another district in Maryland prior to insertion in the benchmark test.

Second, teacher questionnaires for each grade were developed, and teachers were directed to complete a questionnaire for each student prior to the benchmark test administration. The questionnaires asked in detail about the student's mathematics abilities, and also collected data about salient 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*. In all, X and Y questions were asked to 3rd and 5th grade teachers about each student's abilities, respectfully, and these questionnaire data were subsequently keyed to each of the appropriate test items.

These ratings would become the another estimate student ability, 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 which was used was at a similar level of detail to one used by Schmidt, McKnight, Houang, Wang, Wiley, Cogan and Wolfe (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. Further, the ratings were consistent with how the state of Maryland and the district identified specific instructional objectives in their content standards (name, August 2004, personal communication). This meant that the participating teachers were familiar with how mathematics assessment elements would be specified in the ratings. Finally, the approach was piloted with a few teachers in another district and then refined to be useable and feasible for teachers to differentiate student ability. Note that, as the students completed the tests at a later date, the teachers had no knowledge of how the students would score on the items for the project. This would make the criterion measure an independent source of data.

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.

Third, besides the 'plain language' accommodation of the access-based testlets which were administered to all students participating in the project, three supplemental tools were developed to support the text. These were the test-specific Spanish-English and picture-English word lists, and select manipulatives. For the most part, the lists reflected nouns or easily glossed verbs and were keyed by item. The Spanish-English list included a one word or short phrase translation but no definitions. The picture-word list illustrated the relevant English word or phrase.

Select manipulatives were also provided to students identified by the teachers as needing additional tactile or kinesthetic support. For the most part, the manipulatives used in this project were general tools which could be used for several questions, for instance coins or inexpensive counting rods. In all, there were about five manipulatives per grade, and they were placed in a small bag and given to identified students before test time. They could be used as the students needed them for any of the 19 items. The logic behind the use of these tools is explained in Kopriva, 2007. Essentially, while mathematics experts and test designers have generally used manipulatives only to support selected more-complex constructed response items, ELL specialists routinely use tactile and kinesthetic support to provide additional avenues of access in their instruction and classroom evaluation materials. The manipulatives used here were developed or collected to mimic the processes used by these teachers in their classrooms or pullout resource rooms.

Finally, a series of recruiting, training, and test administration guidance materials were developed as part of the project. These were used with temporary staff who were hired to implement the test and accommodations administration for the project. Language liaisons were given additional training to learn about the constraints of their task.

3. Procedures

Three administration accommodations were identified to support the plain language text and language tools. These were small group, oral administration in English, and access to a trained Spanish-English language liaison. Because of agreements with district personnel, students who received any supplementary tools were separated out from the general population and monitored by project staff. As such, formal small group

assignment (as per our algorithms) became, for the most part, a moot point with most beginner and many intermediate students. A relatively small number of advanced students received this accommodation only if teachers thought they needed the intervention.

Reading the mathematics test aloud in English was allowed by the district and was administered to identified ELL students. Further, a small number of bilingual language liaisons were trained and used to aid low literacy Spanish speakers. The role of these liaisons was to translate phrases and words "on the fly" not found in the Spanish-English word list. This was currently not allowed by the state or district, but was allowed for the project and provided at this level to help students with little or no literacy in either language. Unfortunately, low literacy students of other languages did not have the same advantage of either the liaison or the bilingual word list, but the decision was made to go ahead with the Spanish materials as a great majority of the ELLs spoke Spanish as their home language (X%). The accommodation currently used by the district for ELLs was the oral administration in English and this (as well as picture-word list) was still used for lower literacy students whose first language was other than Spanish. Both the oral English and language liaison accommodations were taped and monitored by project staff to minimize the occurrence of cuing or other confounding behaviors.

The agreement with the district is that no students with IEPs would be part of the project. However, native English speaking students who had low reading scores were provided oral administration in English as part of the project. It appeared that a number of these students may have had learning disabilities, but this was not formally acknowledged.

Student questionnaires were sent to teachers to fill out in fall 2004, and school records data which identified students by proficiency level were received at the same time. After records data and completed questionnaires for each student were received by study staff, the ancillary information, as well as data about time in country, proficiency level, and primary language, were used to assign students to appropriate tools and administration accommodations, based on an early version of what would become the STELLA matching algorithms. 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.

To implement the test administration and provide necessary accommodations for the project, staff were hired and trained. Because of the short administration time window (approximately 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, X number of temporary personnel were finally hired to take part in the study and administer the test and the accommodations. The test administration took place in December, 2007.

The administration of the mathematics test 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 the students not identified as needing accommodations for this project, and accommodations for students with IEPs were implemented by the district. For the accommodations which were part of this study, administration was implemented by temporary personnel and monitored by study staff. Participating teachers who filled out questionnaires about each student and temporary staff were paid to complete their work for this project.

4. Planned Analyses

The design for this part of the project included two sets of analyses. To address question's 1 and 2, regression analyses were completed. Question 3 was addressed with a series of critical ratio tests analyzing the differences in conditional probability distributions for different groups of students..

4.1 Regression Analyses

The first set of tests 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. While the ρ_{xy} or R_{xy} as an estimate (and ρ_{xy}^2 / R_{xy}^2 for the squared correlations) are the typical coefficients which researchers use to estimate the relationship between two variables, we believe that both this indicator and the regression coefficient of the target criterion (the beta or β) provide meaningful information. Specifically, we

were interested in both the ability of the test scores to differentiate student ability, as defined by the target ratings (identified by the estimate of β or the regression slope in the relationship), and the amount of predictive variation in the relationship (as identified by the estimate of ρ_{xy}^2). For this analysis, the test score was the dependent variable and the criterion rating was the independent variable.

We expected a reasonable estimate of ρ_{xy}^2 with a reasonably large beta for each group and we anticipated that these results would be similar across groups. As Equation 1 illustrates, ρ_{xy}^2 is actually a composite indicator of both the slope of the relationship and the variation (σ_e). A reasonable ρ_{xy}^2 could be the result of a reasonable β with relatively little variation around the line. However, a smaller ρ_{xy}^2 could also include a reasonable β if the variance (σ_e) around the line is large. Conversely, a larger ρ_{xy}^2 could occur if both the slope and the σ_e are small, and a smaller ρ_{xy}^2 would also occur if the β is small and the σ_e is large. While producing a larger ρ_{xy}^2, it does not seem that a small β and small variation would be very indicative of a useful relationship.

Equation 1 $\rho_{xy} = \mathrm{Cov}(x,y)/((\mathrm{SD}(x))(\mathrm{SD}(y)))$

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

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

$\rho_{xy}^2 = \beta^2 \sigma_x^2/(\beta^2 \sigma_x^2 + \sigma_e^2)$

$R_{xy}^2 = \sim b^2 s_x^2/(b^2 s_x^2 + s_e^2)$

The significance tests for the contrasts were completed by performing Analysis of Covariance tests 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?

4.1 Comparison of Conditional Probability Distributions

To the extent that the regression results were not similar between control and ELL groups, two additional sets of analyses were designed to address research question 3. The analytical question tested in both sets of tests would be: Are there differences between ELL and control groups in their conditional probabilities? Given the students whose teachers indicated that they had some mathematics ability, the probabilities which will be compared are the proportion of those students who receive a testlet score at chance level or below in multiple choice or a similar level in constructed response. Critical-ratio tests of the difference in conditional probabilities would be used to compare the groups, and chi-squares (the square of the critical ratio) would be used to calculate the significance levels.

The first set of analyses would inspect the comparison of the relevant conditional probabilities. All of the students in the focal quadrant would be represented in these analyses, for each grade and subtest.

The second series of analyses would focus on the same conditional probabilities, but examine how particular ancillary variables might continue to impact the control versus the ELL groups. For the purposes of this set of analyses, only those students whose teachers indicated they had problems with the identified variable would be included. This would again limit the number of students in these analyses. If the sample

size for certain ELL groups becomes too small, a composite group of ELLs may have to be constituted.

To determine the relevant data for these analyses, the score x criterion distributions were to be separated into four quadrants (see Figures 1-10 under results below). Since the focus of the inspection was to view the test score/rating relationship for students who teachers said they knew some mathematics, but where their score did not reflect this knowledge, the teacher rating or criterion 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 criterion bar on the other hand was arbitrarily set at 1/3 of the total possible criterion ratings. This was done so that the intended quadrant would underestimate the students who knew some mathematics, increasing the probability that the quadrant was reflecting true score mathematics ability. We reasoned that by using these cut points, fewer possible classifications in the LH quadrant classification (low on scores, mid-high on ratings) would turn out to be false negatives. In other words, if significant differences between the conditional probabilities for the ELL levels vs. native speakers were found, these results would tend to underestimate the problem rather than overestimate it. We were 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 low using both measures (LL) or inconsistently classified (HL). We would also perhaps under-representing the number of students in the HH group. However, these

quadrants weren't the target of the inquiry. Rather, the focus of this set of analyses was to be able to meaningfully interpret the results from the focal (LH) quadrant.

Table 2 presents the cut scores for both indices. Criterion teacher ratings ranged from 1 to 3 per item along a continuum as 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, and total points within each subtest were used. Because of rounding, the cut-points were the same for both grades.

Table 2. Cut Points for Quadrants in Grades 3 and 5

Item Type	*Measure*	*Cut*
Multiple Choice	Criterion	<1.7
	Subtest Scores	<3
Constructed Response	Criterion	<1.7
	Subtest Scores	<4

For the second set of analyses under research question 3, the impact of reading would be the first ancillary variable to be evaluated. The impact of the context, testwiseness, and psychosocial variables will also be evaluated as they were ancillary variables the project staff tried to minimize when the project materials were developed and administration accommodations were selected. Data on each of these student indicators are collected as part of the teacher questionnaires for each student. The questions about reading level of the students asked teachers to rate the student's ongoing reading proficiency in the classroom one a 1-5 scale with 1 indicating reading that is consistently below grade level, and 5, reading consistently above grade level. These analyses would select students at levels 1 and 2, 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 for particular students, those students will included in these analyses. 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 this second set of analyses.

Results

Table 3 presents the mean and standard deviations of the test score results by grade, group, and item type. As expected, for both grades the mean test scores are higher for exited ELLs and native English speakers than for the three ELL groups, and within ELLs, variability in the test scores increase as students gain more English proficiency. Note that the scores of the exited students in grade 3 were higher than native English speakers for both multiple choice and constructed response subtests, and higher for the constructed response subtest for grade 5.

Table 4 present mean and standard deviation data for the teacher ratings. Like the testlet scores, the ratings also reflects an increase in estimated mathematics ability of the students as language proficiency increases. While there seemed to be some more

variability in skills measured the constructed response items in grade 3, as compared to multiple choice, and between native speakers and ELLs at grade 5, all in all, the variability in the target criterion ratings seems to remain similar across groups. 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 ability for ELLs at a similar level of specificity as the specificity identified for native English speakers and exited ELLs.

Table 3: 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)
Native English	5.899 (2.554)	5.752 (3.748)	5.809 (2.705)	6.752 (4.169)
	N = 1245		N = 1175	

Table 4: 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)
Native English	2.291 (0.499)	2.332 (0.509)	2.214 (0.572)	2.160 (0.613)
	N = 1245		N = 1175	

1. Regressions

To address questions 1 and 2, regressions were completed for both grades, using the multiple choice or constructed response subtest scores as the dependent variable, and the criterion measure as the independent variable. Tables below separate out the results for the three proficiency levels, exited ELLs and native English speakers. Post hoc tests were completed which analyzed the difference in slopes for all pairwise group contrasts.

1.1 Grade 3 Regressions

Within each item type, the grade 3 R_{xy}^2 and beta estimate regression results for each group, as well as an omnibus F test, are presented in Table 5. The omnibus F is the main effect of the difference in the beta slopes across groups for each testlet. Contrasts of each set of raw betas are reported in Table 6 where the F-ratios are the interaction Fs in Analysis of Covariance significance tests. In Table 6 the contrasted groups are numbered rather than identified by name. Beginner refers to group 1, Intermediate is group 2,

Advanced is group 2, exited is group 4 and native English speakers is group 5. The contrast results answer the question: Did the covariate interact differently for the first group as compared to the second? Therefore, if the relationship is similar for the two contrasted groups, there would *not* be a significant interaction result.

As the findings in Table 5 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 native English speakers as compared to their English language learner 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 6 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 standard errors), but since the R^2 is also high, the result is promising.

Table 5. Grade 3 Regression Results

		IV	b	se	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			
	Native English	(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		
		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	8.670	0.000
		Target	1.739	0.694	0.015			
	Exited	(Constant)	-2.592	1.162	0.027	0.202		
		Target	3.679	0.469	0.000			
	Native English	(Constant)	-2.565	0.578	0.000	0.234		
		Target	3.568	0.242	0.000			

Table 6. 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	

2.2 Grade 5 Regressions

For Grade 5, results are presented in Table 7 and Table 8. As the findings in Table 7 indicate, the equivalence of the R^2 and betas for this grade generally show more potential. 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. However, for both item types, the results from the

intermediate students are still different from both the 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.

Table 8 confirms 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. Further, the beginners, again, are not significantly different either of the control groups for the constructed response items.

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 7. Grade 5 Regression Results

		IV	b	se	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			
	Native English	(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			
	Native English	(Constant)	-0.142	0.504	0.778	0.220		
		Target	3.192	0.224	0.000			

Table 8. 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	

2. Conditional Probability Analyses

Because of the generally discouraging results for grade 3 and the more mixed but somewhat promising results for grade 5, research question 3 (What is the extent of the problem?) became all the more relevant. As noted above, two sets of analyses were conducted. The specific analytic question to be tested in both sets was: Are there differences between ELL and control groups in their conditional probabilities? The

conditional probabilities are, given the students whose teachers indicated that they had some mathematics ability, the proportion of those students who receive a testlet score at chance level or below in multiple choice or a similar level in constructed response. Critical-ratio tests of the difference in conditional probabilities were used to compare the groups, and chi-squares (the square of the critical ratio) were used to calculate the significance levels.

The first set of tests compared the conditional probabilities of each of the ELL groups (beginner, intermediate, advanced) and exited to the native English speakers. The second examined how particular ancillary variables (reading, testwiseness, context, and psychosocial) impacted the conditional probabilities of the control versus ELL groups. As anticipated, the n's for the beginner and advanced ELLs became too small when the data for the second set of tests were examined, therefore, the three ELL groups were collapsed into a single English language learner group.

2.1 Comparison of Conditional Probabilities for ELL groups/Exited vs. Natives

2.1.1 Grade 3

Figures 1 through 10 illustrate the distribution of test score/teacher rating results graphically for each group (Figures 1-5 depict the multiple choice results and 6-10 depict the distributions of the constructed response test results/criterion ratings). The criterion 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 criterion ratings, respectively, while the diagonal line illustrates the slope of the solution. In the graphs, dots are at each score point reflect

scores/ratings for one or more students (usually multiple students are represented). In each graph the lower left (low test score, low teacher rating (LL)) and upper right (mid-high test score, mid-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 mid-high teacher rating) and this quadrant is found on the bottom right hand side of each graph.

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 for 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 omnibus chi-square result is 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 from both item types, Table 10 reports the critical ratio results with chi square tests of significance for each ELL/native speaker dyad. 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							

2.1.2 Grade5

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 frequency and percentage distribution data by quadrant for multiple choice and constructed results. Like grade 3, very significant omnibus 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) indicate the distributions over quadrants differ 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 ratios/chi square tests 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, where exited students are significantly less apt to be misclassified L/H as compared to their native English peers.

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							

2.2 Comparison of Conditional Probabilities as Further Conditioned on Ancillary Variables

The following four sections will present conditional probability results for students whose teachers said they were consistently or sometimes below grade level reading in the classroom; for students whose teachers said they had testwiseness problems or that they had generally had trouble accessing contexts used in test items; and students whose teachers rated that they observed them exhibiting at least three of five psychosocial problems during test times. Because of sample size ELL groups are collapsed in the critical ratio tests. In the reading section, for each grade and testlet, descriptive tables (along with the critical ratio tables) will be presented for all quadrants and separately for beginners, intermediate, advanced, exited and native speakers. For readability purposes however, only frequency data associated with the relevant conditional probabilities for ELLs (as a collapsed group), exited and native students will be presented within the text for the other analyses.

2.2.1 Reading

This section first presents the descriptive results, by group and quadrant (Table 13) 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 all tables, the n's differ across item type as teacher rating data were collected separately for each item. For the multiple choice conditional probability results, chi-square findings indicate that significantly more ELL students as compared to native speakers score at or below chance when their teachers say they have demonstrated some mathematics ability in the classrooms. There are not significant differences in the conditional probability for ELLs vs. exited students. The significance tests for the constructed response results report

that there are significantly more ELLs than either exited or native speakers whose teachers say they have some mathematics knowledge but score very low on the testlet.

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

Tables 15 and 16 report the fifth grade quadrant and critical ratio results for students whose teachers said are consistently or sometimes below reading proficiency. As with grade 3, the multiple choice conditional probability results indicate that there is a significantly higher proportion of ELL students with reading difficulties who have been misclassified L/H as compared to either exited or native speakers. For the constructed response testlet, no significant difference was found in conditional probabilities between the ELLs and native speakers, but there is between ELLs and exited students.

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

2.2.2 Context

Tables 17 and 18 display critical ratio results for grade 3 and grade 5, respectively. In the tables, the n's for each group's conditional probability are in parentheses, where the denominator is the number of students who the teachers said knew some mathematics, and the numerator is the number of students scoring at or below chance (or the equivalent in constructed response). Overall, the numbers of students whose teachers said they know mathematics and also had problems with the context of items is very small, especially for grade 5. The results indicate that, for both grades, findings parallel 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, conditional probabilities are significantly different for only constructed response in grade 3 but both subtests in grade 5.

Table 17. Context Conditional Probability P(L/H) Comparisons, Grade 3							
		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat (21/134) vs ELL (27/103)	-0.105	0.054	-1.970	3.880	1	0.049
	Exit (9/54) vs ELL (27/103)	-0.095	0.067	-1.431	2.048	1	0.152
CR	Nat (37/105) vs ELL (42/64)	-0.304	0.075	-4.025	16.204	1	0.000
	Exit (10/41) vs ELL (42/64)	-0.412	0.090	-4.604	21.194	1	0.000

Table 18. Context Conditional Probability P(L/H) Comparisons, Grade 5							
		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat (11/84) vs ELL (9/18)	-0.369	0.123	-2.989	8.935	1	0.003
	Exit (2/33) vs ELL (9/18)	-0.439	0.125	-3.516	12.365	1	0.000
CR	Nat (17/74) vs ELL (5/13)	-0.155	0.144	-1.079	1.165	1	0.281
	Exit (2/27) vs ELL (5/13)	-0.311	0.144	-2.156	4.648	1	0.031

2.2.3 Testwiseness

Tables 19 and 20 display results for grade 3 and grade 5. As with context, the probability distributions in both grades indicate that samples of students who are in the L/H condition are quite small, especially ELLs in grade 5. It appears that the only significant differences in the comparison of the conditional probabilities occur for the constructed response subtest in grade 3. All other comparisons, including all in grade 5, suggest that there are no differences between the percentages of ELLs and exited or native speakers who scored low on the testlets and were ranked higher in mathematics ability as estimated by the teachers.

Table 19. Testwiseness Conditional Probability P(L/H) Comparisons, Grade 3							
		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat (10/72) vs ELL (11/48)	-0.090	0.073	-1.235	1.526	1	0.217
	Exit (6/48) vs ELL (11/48)	-0.053	0.089	-0.591	0.349	1	0.555
CR	Nat (24/62) vs ELL (22/31)	-0.323	0.102	-3.152	9.936	1	0.002
	Exit (6/26) vs ELL (22/31)	-0.479	0.116	-4.126	17.022	1	0.000

Table 20. Testwiseness Conditional Probability P(L/H) Comparisons, Grade 5							
		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat (9/42) vs ELL (4/9)	-0.230	0.177	-1.298	1.685	1	0.194
	Exit (2/17) vs ELL (4/9)	-0.327	0.183	-1.784	3.184	1	0.074
CR	Nat (9/39) vs ELL (1/5)	0.031	0.191	0.161	0.026	1	0.872
	Exit (1/13) vs ELL (1/5)	-0.123	0.194	-0.636	0.404	1	0.525

2.2.4 Psychosocial

The findings for the psychosocial variable are the same as those for testwiseness in both grades (see Tables 21 and 22 for grade 3 and grade 5, respectively). That is, the only significant differences in the comparison of the probabilities between ELL and both control groups occur for the constructed response subtest in grade 3. Sample sizes, while still small, are somewhat larger than for testwiseness, however.

Table 21. Psychosocial Conditional Probability P(L/H) Comparisons, Grade 3							
		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat (16/91) vs ELL (18/71)	-0.078	0.065	-1.191	1.418	1	0.234
	Exit (4/34) vs ELL (18/71)	-0.136	0.076	-1.797	3.228	1	0.072
CR	Nat (27/69)vs ELL (27/45)	-0.209	0.094	-2.227	4.958	1	0.026
	Exit (6/26) vs ELL (27/45)	-0.369	0.110	-3.348	11.211	1	0.001

Table 22. Psychosocial Conditional Probability P(L/H) Comparisons, Grade 5

		diff	se diff	critical ratio	Chi-square	df	p
Multiple Choice	Nat (10/92) vs ELL (5/21)	-0.129	0.098	-1.314	1.728	1	0.189
	Exit (2/28) vs ELL (5/21)	-0.167	0.105	-1.589	2.524	1	0.112
CR	Nat (19/82) vs ELL (4/16)	-0.018	0.118	-0.155	0.024	1	0.877
	Exit (4/23) vs ELL (4/16)	-0.076	0.134	-0.568	0.322	1	0.570

Discussion

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

First, overall the quadrant analyses reflect that significantly more beginner and intermediate ELL students as compared to the native English speakers are scoring at chance or below (or the equivalent in constructed response) even though their teachers said they knew some mathematics. This confirms the poor multiple choice regression results for beginners and the multiple choice and constructed response regression results for intermediate students, and implies that the error may not be distributed similarly across the range of scores for different groups. It also suggests that, while the constructed response regression results are promising for beginners at both grades, the errors in these regression findings include significantly more beginner ELLs than native speakers whose teachers said they had some mathematics ability but this ability was not reflected in their test scores. On the other hand, even when the regression results did not show high levels of concurrence between advanced students and native speakers, the conditional probabilities were generally more similar across groups in both grades and for both item types. All in all, it is ELL students who have demonstrated ability elsewhere that accommodations were developed to help. It appears that more work is still needed to

understand how to properly measure the knowledge and skills of especially ELL students with lower levels of English proficiency.

Second, it is important to consider that the amount of non-parity still evidenced in the regression results could occur for a few different reasons. Among them seem to be:

a) Teachers may not be able to estimate the mathematics abilities of lower literacy English language learners (as defined in the teacher questionnaire items) as well as they can for native English speakers and exited students.
b) Teachers may be generally correct in their estimations but one or more types of opportunity to learn issues (OTL) may be more of a problem for English language learners as compared to their native English and exited peers.
c) 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, this does not appear to be the primary issue. The evidence for this inference comes from the similar variability in the ratings of ability across all the groups, including the three levels of ELL and the two control groups. Teachers generally identified low proficient English learners with less mathematics ability than higher level ELLs or non-ELL peers (which we think is reasonable), and at the same time they were able to also identify some of these lower language proficient students as having mathematics knowledge. All in all, the teachers seemed to be able to discriminate the abilities of students within each of the ELL subgroups, and this discrimination seemed to be similar in variability to how they evaluated the abilities of exited and native English speaker students.

On the other hand, we think there is a good chance that both of the other explanations are interacting with test score performance and these disparities seem to be particularly problematic for beginners and intermediate students. “B” suggests there may be non-equivalent mathematics instructional time in classrooms with large numbers of ELLs. Another possibility is that there may be equivalent instructional time, but more of the time may be used to address previous schooling deficiencies in classrooms of students with less language proficiency (and less proficiency in mathematics overall). Additionally, experience suggests that many teachers of ELLs may resort to using computational problems more often than word problems during instruction and evaluation as a way to communicate with the students who have less literacy, and who might otherwise be faced with language difficulties on word problems if the items are not phrased, presented or scored correctly. This would suggest that these students may have insufficient practice in how to interact with word problems, and this would contrast to the assessment opportunities of most non-ELL students. As one reviews the validity results found in this study, it appears that, in fact, beginner students seem to appreciate context-based items but they need to be given a chance to explain their thinking.

As noted above, “c” considers that the measurement of academic content still needs to be improved for ELLs, particularly those with lower levels of literacy in English. Perhaps the correct accommodations were not used for large numbers of beginner and intermediate students, or perhaps the form and format of how they are tested needs to be re-conceptualized. Future research should consider expanding how to systematically address this population of students with little literacy, identifying some new tools or new

approaches to documenting ability in ways that are comparable across approaches as well as groups.

The third implication of this study reflects the finding that response options for beginner students seem to be an essential accommodation. While the stems or prompts for both the multiple choice and constructed response items were similar in language and format, the abilities of 3rd and 5th grade beginner ELL students (as estimated in the teacher ratings) corresponded poorly to multiple choice test scores but showed promise when measured with constructed response items. Although the L/H conditional probability for the constructed response scores still indicates a significantly greater misclassification proportion for beginners as compared to native speakers and exited, the distance between the groups seems to be the narrowest when these students are allowed flexibility in how to respond.

Conversely, the multiple choice results seem to underscore the language and cultural issues which exist in the format of multiple choice items but are frequently overlooked by mainstream test developers. This includes cultural assumptions underlying the choosing of options, the language of discrimination which is necessary for making choices, and the increased language load in the answer options. It is important to understand that the multiple choice item type, as an approach to asking questions, does not have students explain per se, but has them discriminate among available options and make an informed choice. Because many ELLs come with previous formal and informal learning experiences that are quite distinct from mainstream non-ELLs, the choices are often not considered in the same way as they are for native English speakers.

Further, the language of discrimination is not naturally words and phrases that a student new to a language (in this case English) would learn. It is possible that minimizing the ways discrimination questions are worded may help, for instance, if standard question forms are taught to students. But it is sometimes impossible to use these forms to properly convey the meaning in the item. Visuals in the options may be a partial solution to the language load issue, but the choices in many multiple choice items don't lend themselves easily or effectively to a static illustration, especially when students are asked to discriminate fairly finely among the options.

In addition to constructed response, there may be other item types or testing approaches that address some of the challenges in how items are presented to students, and also increase the response options for students who must rely on strengths other than language to convey their knowledge and skills. For instance, it may be time to consider paper and pencil items that use other close ended formats, or computer-based items which can simulate context without resorting to language, and allow students to interact with stimuli to demonstrate their responses.

A fourth consideration from the study findings are the consistent results from intermediate English language learners, over item types and grades. 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 English speakers who know how to gloss over these inconsequential textual elements. This fact may partially explain

why intermediate students at both grades do not seem to respond as well as exited/native English speaking students.

Hyper-sensitivity may also partially explain the results for the advanced ELLs at grade 3. In third grade, when most students are still perfecting their basic reading skills, the advanced ELLs, may be facing both the challenges associated with literacy along with their native speaking peers, and also the language limitations of being a young ELL. As a consequence they may temporarily "look" like intermediate ELLs when asked to perform on tests, using discrimination and evaluative skills that may not be on par with other English proficiency skills they demonstrate in less pressured situations. This could be true especially in constructed response where the quadrant analyses showed a significantly greater proportion of 3rd grade advanced students were misclassified by these test scores as compared to native English speakers. On the other hand, since the quadrant analyses for the advanced group at grade 3 did not find that there appeared to be a misclassification of ability and multiple choice test scores, there seems to also be some other reason why these students do not navigate this type of item well at this grade level. Either or both of these reasons may explain the discrepancies between teacher ratings and test scores for this population.

The convergence of developmental skills and literacy skills at grade 3 may point to why advanced ELLs seem to be a transition group, especially between grades 3 and 5. As all students mature developmentally, they learn how to construct symbolic frameworks to cope with their schooling and living. Even though advanced ELLs at grade 5 still have language and literacy challenges, as part of their developmental maturity they may have developed other compensatory skills to partially offset the ELL challenges. It

seems that perhaps the skills of both handling textual inconsistencies, and navigating the language, text, and schooling expectations, impacts test performance for advanced students differently at two different grade levels.

Fifth, consistently, reading as an ancillary variable still remains an issue for English language learners in these universally designed items. Context also appears to affect the younger students more often than older students. This may be a result of less experience in U.S. schools, and less experience in general with U.S. culture and experiences. However, while n's are small, context for older students, as well as psychosocial variables, continue to affect the conditional probability comparisons significantly more often for ELLs as compared to native or exited students. Testwiseness was the one variable that, on the whole, seemed to be leveled in its effect on scores across groups. The only exception to this is in the 3rd grade constructed response subtest, where the conditional probabilities of ELL students are significantly higher than the probabilities for native English speakers and exited.

All in all, the L/H probability sample sizes for context, psychosocial and testwiseness tend to be small, though they are consistently larger for third grade than fifth. Substantial numbers of students are still impacted by reading, and the n for context is considerable for grade 3. This suggests that context, psychosocial and testwiseness variables might tend to interact with reading and/or developmental issues more at the younger grades and become less a factor as students mature. Further, item type differences in relation to the variables need to be considered. When one observes that some students from all groups continue to be impacted by the ancillary factors identified here, there is a reasonably large group of students in most large-scale test taking

situations who could benefit from additional improvements. Ancillary variables such as these suggest a few points of focus for doing this work. Some of these variables may aid various students with language difficulties, while some will undoubtedly be associated with unique ELL variables which affect language acquisition and cultural experiences.

Finally, the conditional probability results for the constructed response subtests indicate that, consistently, there are significantly more Exited students than native English speakers who have ability in mathematics and receive test scores higher than the equivalent of chance. 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. While there is no significant differences in the betas across these two groups, the R^2 coefficients in the regression results for exited students are more similar to those of native English speakers for the constructed response subtests in both grades (with higher coefficients in grade 5), as compared the results for multiple choice scores. Thus, both the regression and conditional probability findings suggest that exited students benefit from the open-ended quality of this item type which allows them to explain what they know. Language may still be an issue for these students, even though they "do all right" with the multiple choice format.

Initially, these findings were greeted with frustration, confusion and concern. After all, researchers and interested practitioners have been working hard for more than the last decade to improve large-scale testing for English learners. How could results be this poor? Upon reflection, however, it seems that the work completed to date has been useful, to minimize barriers for more advanced ELLs, and to provide guidance as to how to proceed from here. Multiple choice tests have hidden problems for lower literacy

ELLs, and this population, in particular, seems to need response accommodations as well as presentation and administration options. We don't seem to be meeting the needs of intermediate students yet, and they seem to have problems both similar to and distinct from their beginner ELL peers that need to be understood. The most well developed access-based items are helpful for a number of ELLs, to a point, but static illustrations and text necessary to convey complexity are limits that need to be overcome. Reading (or whatever it might be a surrogate for) is still an issue, and, when developing text and assigning accommodations, one needs to consider age as well as English language proficiency. So, all in all, this study seems to serve as a status check. There have been improvements, but more needs to be done. It is hoped that the results will help researchers and practitioners continue to design studies that add to the knowledge gleaned to date, and work to continue to improve testing for all English language learners.

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.

Rivera, C., Collum, E., Shafer Willner, L, & Sia, J. (2006). Analysis of State Assessment Policies Regarding the Accommodation of English Language Learners in C. Rivera & E. Collum (Eds.) State Assessment Policy and Practice for English Language Learners: A National Perspective. Lawrence Erlbaum Associates, Inc.