Evaluating Overall Proficient Level Cut on Arizona English Language Learner Assessment Using a Parametric Regression Discontinuity Design with a Binding Score

Stages III through V

February, 2016

Prepared by Pearson for Arizona Department of Education

Executive Summary

The Overall Proficient cut for the Arizona English Language Learner Assessment (AZELLA), which was set at the standard setting meeting in 2013, was evaluated for Stages III through V using a regression discontinuity design. The regression discontinuity design is a quasi-experiment design used to evaluate the appropriateness of the reclassification of ELL students as Reclassified Fluent English Proficient (RFEP). The regression discontinuity is a powerful tool for cases when investigating such effects is *not* practically feasible under a randomized experiment design (Lee & Munk, 2008), which is the case with ELL reclassification. Robinson (2011) argued that if a threshold to reclassify ELL students is appropriately set, there should be no effect at the threshold to make the transition from an ELL program to the mainstream classrooms. To investigate the performance of ELL and RFEP students on Arizona's Instrument to Measure Standards (AIMS) Reading, Mathematics, and Writing in the following spring, a study was implemented using the parametric regression discontinuity design approach by a regression analysis with a binding score of Spring 2013 AZELLA scores for Total Combined, Reading, and Writing. The results from the analysis suggested that the AZELLA Overall Proficient cut was appropriately set for ELL reclassification for most of the grade and subject combinations investigated in the study.

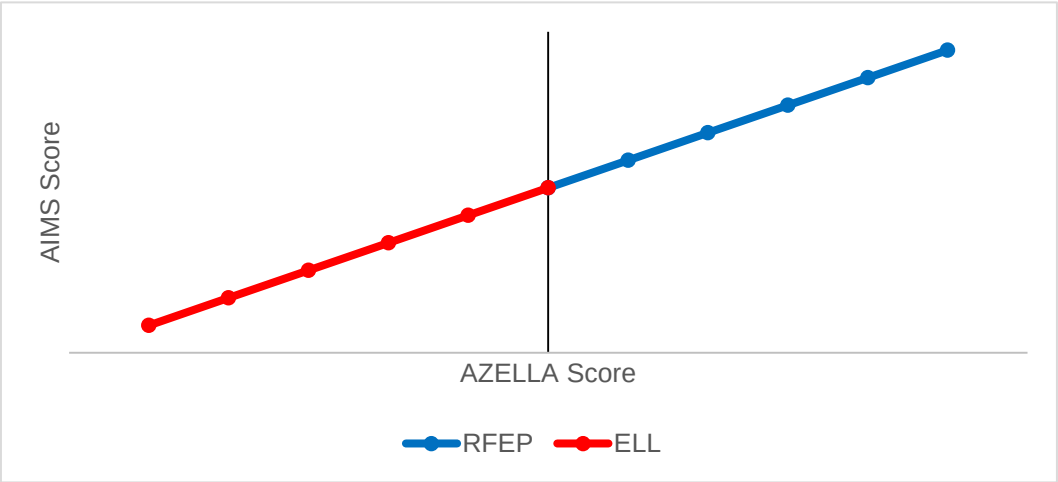
Introduction

AZELLA measures English Proficiency based on the current Arizona ELP Standards. AZELLA includes its Placement and Reassessment tests for five stages (Stage I: Kindergarten, Stage II: Grades 1 and 2, Stage III: Grades 3 through 5, Stage IV: Grades 6 through 8, and Stage V: Grades 9 through 12). The Placement tests for Stages II through V and the Reassessment tests for Stages I through V are vertically scaled and equated. (Please note that the new Kindergarten Placement Test, which is used as a screener, is not on the same scale as the AZELLA test.) The proficiency standards for the most recent version of AZELLA were set at the standard setting meeting in 2013 (Arizona Department of Education, 2013). The Total Combined score, which is a composite of the Listening, Speaking, Reading, and Writing scores, has four proficiency levels (Pre-emergent/Emergent, Basic, Intermediate, and Proficient). In contrast, each of the other domains and subdomains (Listening, Speaking, Reading, Writing, Language, Oral Communication, Comprehension, and Literacy) has only three proficiency levels (Pre-emergent/Emergent/Basic, Intermediate, and Proficient). The AZELLA Overall Proficient level is conjunctive, not compensatory; specifically, students must earn Proficient on each of Total Combined, Reading, and Writing in order to be Overall Proficient. ELL students in Arizona who earn Overall Proficient on the Reassessment in spring exit the ELL program and make the transition to mainstream classrooms as RFEP the next school year. The other students who do not earn Overall Proficient on the Reassessment in spring will remain as ELL students.

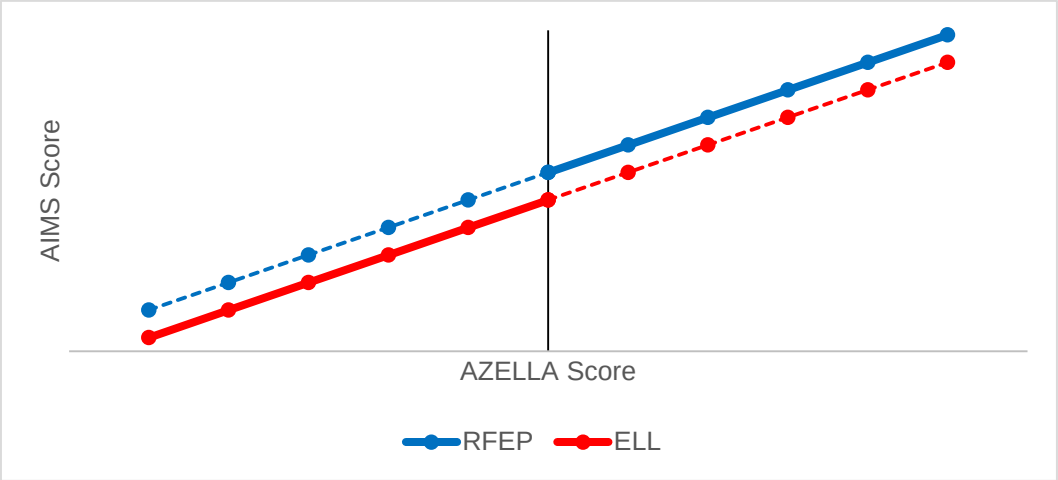
The objective of this study was to evaluate the Overall Proficient cut for the Stage III through V by performing a regression discontinuity design (Jacob & Zhu, 2012; Lee & Munk, 2008; Robinson, 2011; Smith, 2014). The regression discontinuity design is a quasi-experiment design used to evaluate the appropriateness of the reclassification of ELL students as RFEP. The regression discontinuity is a powerful tool for cases when investigating such effects is *not* practically feasible under a randomized experiment design (Lee & Munk, 2008).

Robinson (2011) argued that if a threshold (e.g., AZELLA Overall Proficient cut) to reclassify ELL students is appropriately set, there should be no effect at the threshold to make the transition from an ELL program to the mainstream classrooms. The argument can be explained by hypothetical examples in Figure 1. In three panels of Figure 1, it is assumed that AIMS scores and AZELLA scores have a linear relationship, and RFEP students are dummy coded as 1 and ELL students are dummy coded as 0. A vertical line in each panel represents an AZELLA cut score so that RFEP students are at the right side of the AZELLA cut score while ELL students are at the left side of the cut score. Solid lines in red and blue in three panels represent observed AIMS scores given an AZELLA score for ELL and RFEP students, respectively. A dotted line in blue represents expected AIMS scores, given AZELLA scores, if ELL students were reclassified as RFEP and went to the mainstream classrooms; whereas, a dotted line in red shows expected AIMS scores given AZELLA scores if RFEP students stayed in an ELL program. Panel a shows ‘no effect’ of ELL reclassification as a desirable outcome because there is no disconnect between regression lines for RFEP and ELL students. This implies that the transition from the ELL program to the mainstream classrooms for students is appropriate. In other words, the AZELLA cut score is appropriate for ELL reclassification. Panel b shows ‘positive effects’ because the AIMS score is higher for RFEP students than that for ELL students at the AZELLA cut score. In Panel b, the expected AIMS scores for ELL students (if ELL students were reclassified as RFEP) is higher than the observed AIMS scores. This indicates that ELL students are held in the ELL program longer than they should. In other words, ELL students could get more benefits in the mainstream classrooms. This implies that the AZELLA cut score is higher than it should for ELL reclassification. In contrast, Panel c represents ‘negative effects’ because the AIMS score is lower for RFEP students than that for ELL students at the AZELLA cut score. The expected AIMS scores for RFEP students (if RFEP students were retained as ELL) is higher than the observed AIMS scores. This means that RFEP students exit the ELL program earlier than they should and the AZELLA cut score is lower than it should.

(a)



(b)



(c)

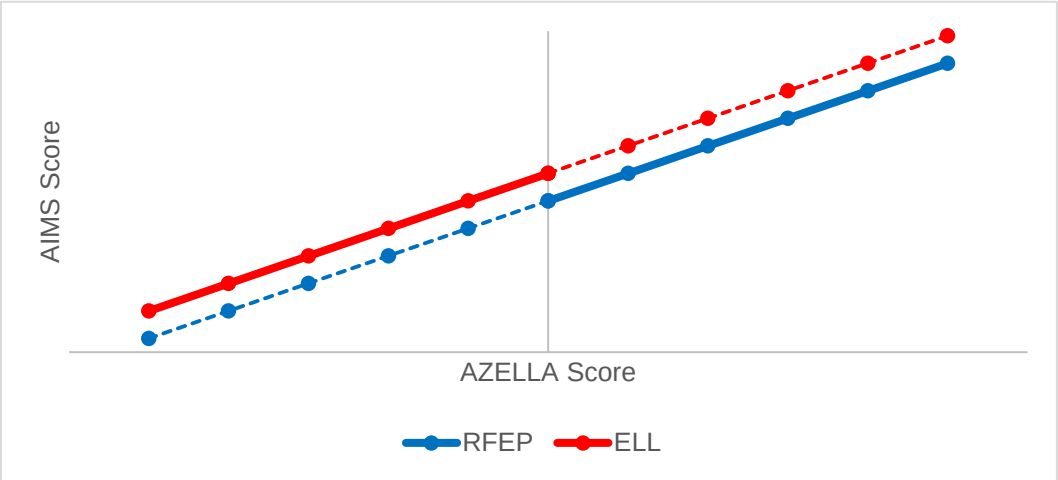


Figure 1. *Graphical illustration of different effects in regression discontinuity design.*

A parametric model for the regression discontinuity design (Jacob & Zhu, 2012; Lee & Munk, 2008; Robinson, 2011; Smith, 2014) is used in this study.

A Parametric Regression Discontinuity Design with a Binding Score

A parametric regression discontinuity design generally involves a regression analysis with a single cut point or threshold. On the other hand, AZELLA uses multiple thresholds (i.e., the Proficient cuts on Total Combined, Reading, and Writing) to reclassify ELL students. Robinson (2011) proposed that a binding score should be constructed to reduce the dimensionality of multiple cut scores to a single score. The binding score for this study is defined as the minimum of AZELLA Total Combined, Reading, Writing scale scores standardized and recentered around their respective Proficient cut score, or formally $M_j = min$(AZELLA Total Combined scale score, AZELLA Reading scale score, AZELLA Writing scale score), in which these scale scores in the formula are standardized and recentered around their respective Proficient cut. With this approach, ELL students who earn Proficient on AZELLA Total Combined, Reading, and Writing and thus earn Overall Proficient always have binding scores greater than or equal to 0. On the other hand, ELL students who earn below Proficient on at least one of the three criteria always have negative binding scores. In other words, ELL students who earned Overall Proficient and were reclassified as RFEP based on the AZELLA Reassessment in spring 2013 had the positive binding scores while the other ELL students who did not earn Overall Proficient and were retained as ELL after the AZELLA Reassessment in spring 2013 had the negative binding scores.

A general regression model for the regression discontinuity design with the binding score used in this study is the following equation:

$$Y_j = \alpha + \delta R_j + f(M_j) + \varepsilon_j, \quad (1)$$

where Y_j is AIMS standardized scale score for student j, α is an intercept, δ is an effect of ELL reclassification on the AIMS standardized scale score, after controlling for the AZELLA binding score, M_j, R_j is a dummy code for ELL reclassification (i.e., 1=RFEP, 0=ELL), f is a function of M_j, and ε_j is an error term. δ is the parameter of interest in the regression model. As Robinson (2011) argued, if δ is not statistically significant, it indicates that ELL students who passed AZELLA in spring 2013 had a smooth transition to the mainstream classrooms as RFEP and thus the AZELLA Overall Proficient cut was set appropriately. On the other hand, a statistically significant and positive effect of δ suggests that the ELL students who barely failed the test might have gained more benefits from moving to the mainstream classroom as RFEP while a statistically significant and negative effect of δ suggests that the RFEP students who barely passed the test might have gained more benefits from staying in an ELL program. In other words, the positive effect indicates

that the AZELLA Overall Proficient cut might have been set too high; whereas, the negative effect indicates that the cut might have been set too low.

As suggested by Jacob and Zhu (2012), Six regression models were ran in this study to determine the best model among them in order to minimize the bias of estimating the effect, δ. The regression models are as follows:

Linear: $Y_j = \alpha + \delta R_j + \beta_1 M_j + \varepsilon_j,$ (2)

Linear Interaction: $Y_j = \alpha + \delta R_j + \beta_1 M_j + \beta_2 R_j M_j + \varepsilon_j,$ (3)

Quadratic: $Y_j = \alpha + \delta R_j + \beta_1 M_j + \beta_2 M_j^2 + \varepsilon_j,$ (4)

Quadratic Interaction: $Y_j = \alpha + \delta R_j + \beta_1 M_j + \beta_2 M_j^2 + \beta_3 R_j M_j + \beta_4 R_j M_j^2 + \varepsilon_j,$ (5)

Cubic: $Y_j = \alpha + \delta R_j + \beta_1 M_j + \beta_2 M_j^2 + \beta_3 M_j^3 + \varepsilon_j$, and (6)

Cubic Interaction: $Y_j = \alpha + \delta R_j + \beta_1 M_j + \beta_2 M_j^2 + \beta_3 M_j^3 + \beta_4 R_j M_j + \beta_5 R_j M_j^2 + \beta_6 R_j M_j^3 + \varepsilon_j.$ (7)

Residuals from all regression models were compared to determine which model fitted the data best as the final model as suggested by Robinson (2011).

Data

AZELLA Reassessment data from spring 2013 were used for this study. Only students with ELL status at the time of assessment were included in any analysis because the focus of study is to evaluate the reclassification of ELL students. Note that RFEP and non-ELL students were also included in the administration. Those ELL students in the AZELLA data who earned Overall Proficient by meeting the Proficient level on each of the Total Combined, Reading, and Writing scores were reclassified as RFEP. The other ELL students were retained as ELL. The AZELLA data was merged with AIMS Reading, Writing, and Mathematics from spring 2014 by the student identification number, and only students who moved up a grade between spring 2013 and spring 2014 were kept in the merged data. AIMS Reading and Mathematics were administered for Grades 3 through 8 and high school, while AIMS Writing was administered for Grades 5 through 7 and high school. For high school, students in cohort 16, who were mostly at grade 10 were kept in the analysis because the target population of AIMS high school was sophomore and the group was in cohort 16 in spring 2014. The students in cohort 16 are called high school students hereafter. All students in the merged data had valid AZELLA Reassessment scores from spring 2013 and AIMS scores from spring 2014. The grade levels presented throughout the study are henceforth based on AIMS data from spring 2014. The number of students in the merged data for AZELLA/AIMS Reading, Mathematics, and Writing by grade is presented in Table 1.

Table 1: The Number of Students in Merged Data

	Grade						
AIMS	3	4	5	6	7	8	High School
Reading	11551	5939	5757	4486	2874	1427	1139
Mathematics	11553	5937	5756	4486	2873	1426	1139
Writing	-	-	5752	4468	2863	-	1125

After the AZELLA and AIMS data were merged, the binding score (AZELLA binding score) was constructed by taking the minimum of scale scores among AZELLA Total Combined, Reading, Writing, which had been already standardized and recentered around their respective Proficient cut for each student in the merged data. AIMS scale scores for the students in the merged data sets were also standardized after merging.

As Robinson (2011) pointed out, the objective of the study with a regression discontinuity design is to estimate the effect of reclassification for the students just below or above the cut point. Thus, the only students in the merged data with their binding scores between -1 and 1 were included as the analytic sample in any further analysis, which is analogous to Robinson (2011). Descriptive statistics for the analytic sample for AIMS Reading, Mathematics, and Writing are presented in Tables 2 through 4, respectively. In this study, at least 100 students were included in each group (e.g., ELL and RFEP) to conduct the regression discontinuity design as suggested by Smith (2014). Descriptive statistics for demographic information (i.e., gender, Hispanic, free/reduced lunch, special education) were calculated including students with missing information.

Table 2: Descriptive Statistics for the Analytic Sample on AIMS Reading

	Grade						
Variable	3	4	5	6	7	8	High School
N-count	8219	1983	3148	2848	1635	741	493
ELL (%)	57.90	76.50	66.87	62.32	73.52	73.01	70.99
RFEP (%)	42.10	23.50	33.13	37.68	26.48	26.99	29.01
Male (%)	52.03	54.51	52.67	54.42	56.64	57.09	53.55
Female (%)	47.90	45.49	47.30	45.54	43.36	42.91	46.45
Hispanic (%)	89.37	86.18	87.23	86.24	84.95	78.81	67.14
Free/Reduced Lunch (%)	88.32	88.86	88.31	88.66	88.20	85.29	82.56
Special Education (%)	9.48	9.23	11.28	16.64	18.53	22.94	17.04
Average AZELLA Total Score	0.12	-0.10	-0.01	0.10	0.01	-0.01	-0.01
Average AZELLA Reading Score	0.24	-0.03	0.04	0.12	0.00	0.01	0.04
Average AZELLA Writing Score	0.22	0.07	0.15	0.26	0.07	0.03	0.01
Average AZELLA Binding Score	-0.13	-0.34	-0.25	-0.16	-0.29	-0.29	-0.26
Average AIMS Score	0.20	0.63	0.38	0.24	0.25	0.23	0.32

Note: RFEP students were ELL students reclassified after passing AZELLA in Spring 2013.

Table 3: Descriptive Statistics for the Analytics Sample on AIMS Mathematics

	Grade						
Variable	3	4	5	6	7	8	High School
N-count	8222	1983	3148	2847	1634	740	526
ELL (%)	57.89	76.50	66.87	62.35	73.50	73.11	71.67
RFEP (%)	42.11	23.50	33.13	37.65	26.50	26.89	28.33
Male (%)	52.04	54.51	52.67	54.48	56.61	57.03	53.80
Female (%)	47.88	45.49	47.30	45.49	43.39	42.97	46.20
Hispanic (%)	89.37	86.18	87.23	86.20	85.07	78.92	67.68
Free/Reduced Lunch (%)	88.32	88.86	88.31	88.65	88.25	85.27	82.70
Special Education (%)	9.47	9.23	11.28	16.65	18.54	22.84	17.49
Average AZELLA Total Score	0.12	-0.10	-0.01	0.10	0.01	-0.01	-0.03
Average AZELLA Reading Score	0.24	-0.03	0.04	0.12	0.00	0.01	0.01
Average AZELLA Writing Score	0.22	0.07	0.15	0.26	0.07	0.03	0.00
Average AZELLA Binding Score	-0.13	-0.34	-0.25	-0.16	-0.29	-0.29	-0.29
Average AIMS Score	0.15	0.52	0.32	0.17	0.18	0.14	0.25

Note: RFEP students were ELL students reclassified after passing AZELLA in Spring 2013.

Table 4: Descriptive Statistics for the Analytic Sample on AIMS Writing

	Grade			
Variable	5	6	7	High School
N-count	3146	2838	1630	522
ELL (%)	66.91	62.37	73.44	72.22
RFEP (%)	33.09	37.63	26.56	27.78
Male (%)	52.54	54.40	56.63	54.41
Female (%)	47.43	45.56	43.37	45.59
Hispanic (%)	87.22	86.15	85.03	67.24
Free/Reduced Lunch (%)	88.30	88.69	88.22	82.76
Special Education (%)	11.28	16.67	18.47	18.20
Average AZELLA Total	-0.01	0.10	0.01	-0.04
Average AZELLA Reading	0.04	0.12	0.00	0.00
Average AZELLA Writing	0.15	0.26	0.07	0.00
Average AZELLA Binding Score	-0.25	-0.16	-0.29	-0.30
Average AIMS Score	0.34	0.25	0.27	0.42

Note: RFEP students were ELL students reclassified after passing AZELLA in Spring 2013.

Results

In order to determine the best regression model among six models in Equations 2 through 7, the following analyses were conducted. First, a scatter plot of AIMS scores against AZELLA binding scores for the analytic sample on each of grade and subject combination was visually inspected. Second, the effect of ELL reclassification on the AIMS standardized scale score (δ) for all regression models were reviewed. Third, the average residuals were plotted and visually examined. From the evaluation, the linear regression model was determined as the final model for each grade and subject combination. The details of model selection are presented in Appendix A.

The parameter estimates of ELL reclassification on the AIMS standardized scale score (δ) from the Linear regression model for all grade and subject combinations are presented in Table 5. The results showed that all grade and subject combinations, except for Grade 8 Reading and Grade 6 Mathematics, had a statistically

non-significant effect, controlling for the AZELLA binding score. Grade 8 Reading had the statistically significant and positive effect, after controlling for the AZELLA binding scores, while Grade 6 Mathematics had the statistically significant and negative effect, after controlling for the AZELLA binding scores.

The largest effect of ELL reclassification on the AIMS standardized scale score for Reading was 0.25 in the *SD* unit at Grade 8, which is equivalent to 8.5 in AIMS scale score (*SD* in AIMS scale score for the merged dataset was 33.68). In other words, the estimated difference in AIMS Reading scale score between RFEP and ELL students at Grade 8 was 8.5 at the AZELLA binding score cut. The minimum conditional standard error of measurement (CSEM) associated with the scale score for the assessment in Spring 2014 was 15 (Arizona Department of Education, 2014). Thus, the difference of 8.5 in the AIMS Reading scale score fell within 1*SE*. Similarly the largest effect for Mathematics was -0.19 in the *SD* unit and -6.2 in the AIMS scale score (*SD* in AIMS scale score for the merged dataset was 33.14) at Grade 6. The difference of -6.2 also fell well within 1*SE* since the minimum CSEM for the assessment in Spring 2014 was 11 (Arizona Department of Education, 2014). In addition, the largest effect in Writing was – 0.10 in the *SD* unit at High School and -4.0 (*SD* in AIMS scale score for the merged dataset was 40.98). The difference of -4.0 also fell within 1*SE*, given the minimum CSEM for the assessment was 9 (Arizona Department of Education, 2014). These findings indicated that the differences in the AIMS scale score between RFEP and ELL at the AZELLA binding score cut may not be significant from the measurement perspective although some of the estimates were statistically significant at 0.05.

Table 5: Parameter Estimate of ELL Classification

	Grade													
	3		4		5		6		7		8		High School	
Subject	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*
Reading	-0.05	0.03	0.03	0.08	0.01	0.06	-0.01	0.06	-0.05	0.08	*0.25	0.13	0.04	0.17
Mathematics	0.04	0.03	0.16	0.09	0.11	0.06	*-0.19	0.07	-0.11	0.09	0.10	0.14	0.19	0.18
Writing	N/A	N/A	N/A	N/A	-0.03	0.06	-0.04	0.06	-0.08	0.08	N/A	N/A	-0.10	0.12

* Significance level of 0.05

Conclusions

A regression discontinuity design, using a regression analysis with a binding score, was utilized to investigate the effect of ELL reclassification on the performance on AIMS. The study results revealed that the AZELLA Overall Proficient cut to reclassify ELL students as RFEP or retain them as ELL was statistically appropriate for most of the grade and subject combinations included in the study. Across three test subjects and 7

different grades for a total of 18 unique samples, the results for only two grade and subject combinations might show the statistical evidence that called the validity of respective AZELLA Overall Proficient cut into question; however, these effects were in different directions and the effects were trivial from the measurement perspective. For Grade 8 Reading, ELL students might have been held in the ELL program longer than they should have while, for Grade 6 Mathematics, they might have been exited the ELL program earlier than they should have. In addition, the results of those grades for another subject(s) showed that the respective AZELLA Overall Proficient cut was set appropriately.

References

Arizona Department of Education (2013). *Arizona English Language Learner Assessment standard setting report: Stage I – V*. Iowa City, IA: Pearson. Retrieved from http://www.azed.gov/assessment/files/2014/04/az-azella-standard-setting-report-stages-i-v_final_100113a.pdf

Arizona Department of Education (2014). *Arizona's Instrument to Measure Standards 2014 Technical Report*. Iowa City, IA: Pearson.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition.* Hillsdale, N.J.: Lawrence Erlbaum.

Jacob, R. T., Zhu, P., Somers, M. A., & Bloom, H. S. (2012). *A practical guide to regression discontinuity*. New York: MDRC. Retrieved from http://www.mdrc.org/sites/default/files/regression_discontinuity_full.pdf.

Lee, H., & Munk, T. (2008). *Using regression discontinuity design for program evaluation*. Denver, CO: American Statistical Association.

Robinson, J. P. (2011). Evaluating criteria for English learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis*, *33*, 267-292.

Smith, W. C. (2014). Estimating unbiased treatment effects in education using a regression discontinuity design. *Practice Assessment, Research & Evaluation, 19*. Retrieved from http://www.pareonline.net/getvn.asp?v=19&n=9.

Appendix A: Model Selection

A series of analyses were conducted to determine the best model among six regression models considered in this study in order to minimize the bias of estimating the effect, δ.

First, the relationship between AIMS scores and AZELLA binding scores for the analytic sample on each of grade and subject combination was visually examined by drawing a scatter plot of the two variables as presented in Appendix B. The scatter plots with fitted linear regression lines for ELL and RFEP groups seemed to suggest a linear relationship between the AZELLA binding scores and AIMS scores for all grade and subject combinations. No clear discontinuity at the cut point (i.e., the AZELLA binding score = 0) was observed from the scatter plots for any grade and subject. Correlation analysis was conducted for both whole sample in the merged datasets and the analytic sample to examine the strength of association between the AIMS scores and AZELLA binding scores. The correlation coefficients presented in Table A.1 showed that the correlation was higher for whole sample than the analytic sample because the AZELLA binding score had the restricted range from -1 to 1. The correlations for the analytics sample among three subjects ranged from 0.32 to 0.47, which approximately translated to medium to large effect size according to the criteria by Cohen (1988).

Table A.1: Correlation between AIMS Score and AZELLA Binding Score

		Grade						
Subject	Sample	3	4	5	6	7	8	High School
Reading	Whole	0.65	0.57	0.57	0.52	0.54	0.49	0.51
	Analytic	0.47	0.41	0.40	0.39	0.39	0.42	0.42
Mathematics	Whole	0.55	0.49	0.49	0.39	0.44	0.35	0.44
	Analytic	0.40	0.36	0.35	0.31	0.34	0.32	0.32
Writing	Whole	N/A	N/A	0.53	0.52	0.53	N/A	0.63
	Analytic	N/A	N/A	0.34	0.35	0.36	N/A	0.46

Then, the outcomes of all regression models with respect to the parameter estimates of ELL reclassification (δ) on the AIMS standardized scale score, presented in Tables A.2 through A.4, were reviewed for each of grade and subject combination. It revealed that the parameter estimates produced by all regression models for all of grade and subject combinations, except for Grade 8 Reading and Grades 5 through 7 Mathematics, were statistically non-significant at the significance level of 0.05, after controlling for the AZELLA binding scores. Note that a significance level correction (e.g., Bonferroni's correction) was not used to check the

statistical significance of δ from the six regression models so that the significance level was more liberal. These results suggest that, no matter which regression model was chosen, there was no statistical evidence for discontinuity at the cut point for those grade and subject combinations, i.e., these particular cut points were appropriate. Nevertheless, after controlling for the AZELLA binding scores, Grades 8 Reading and Grade 5 Mathematics had a positive effect for the Linear and Cubic Interaction models, respectively, whereas Grade 6 Mathematics had a negative effect for the Linear, Linear Interaction, Quadratic, and Cubic Interaction models, and Grade 7 Mathematics had a negative effect for the Cubic Interaction model.

Table A.2: Parameter Estimate of ELL Classification on AIMS Reading

	Grade													
	3		4		5		6		7		8		High School	
Model	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*
Linear	-0.05	0.03	0.03	0.08	0.01	0.06	-0.01	0.06	-0.05	0.08	*0.25	0.13	0.04	0.17
Linear Interaction	-0.04	0.03	0.03	0.08	0.01	0.06	-0.01	0.06	-0.07	0.09	0.13	0.13	0.03	0.17
Quadratic	-0.04	0.03	0.06	0.08	0.02	0.06	-0.01	0.06	-0.09	0.09	0.11	0.14	0.04	0.18
Quadratic Interaction	-0.05	0.05	0.10	0.12	0.08	0.09	0.01	0.09	0.00	0.14	0.07	0.20	0.07	0.28
Cubic	-0.05	0.04	0.09	0.11	0.06	0.08	0.02	0.08	-0.07	0.11	0.08	0.17	0.04	0.24
Cubic Interaction	-0.08	0.07	0.28	0.18	0.11	0.15	-0.10	0.11	-0.03	0.20	0.25	0.30	0.27	0.40

* Significance level of 0.05

Table A.3: Parameter Estimate of ELL Classification on AIMS Mathematics

	Grade													
	3		4		5		6		7		8		High School	
Model	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*
Linear	0.04	0.03	0.16	0.09	0.11	0.06	*-0.19	0.07	-0.11	0.09	0.10	0.14	0.19	0.18
Linear Interaction	0.03	0.03	0.14	0.09	0.11	0.06	*-0.19	0.07	-0.14	0.10	0.07	0.15	0.14	0.19
Quadratic	0.03	0.04	0.15	0.09	0.09	0.06	*-0.20	0.07	-0.14	0.10	0.05	0.15	0.13	0.19
Quadratic Interaction	-0.03	0.05	0.15	0.13	0.19	0.10	-0.11	0.09	-0.27	0.15	0.19	0.23	0.24	0.30
Cubic	-0.01	0.05	0.15	0.12	0.16	0.08	-0.11	0.08	-0.19	0.12	0.10	0.19	0.21	0.26
Cubic Interaction	-0.10	0.08	0.28	0.20	*0.32	0.16	*-0.24	0.12	*-0.45	0.21	0.55	0.34	0.29	0.43

* Significance level of 0.05

Table A.4: Parameter Estimate of ELL Classification on AIMS Writing

	Grade							
	5		6		7		High School	
Model	δ	*SE*	δ	*SE*	δ	*SE*	δ	*SE*
Linear	-0.03	0.06	-0.04	0.06	-0.08	0.08	-0.10	0.12
Linear Interaction	-0.02	0.06	-0.04	0.06	-0.07	0.08	-0.10	0.13
Quadratic	0.00	0.06	-0.05	0.06	-0.08	0.08	-0.10	0.13
Quadratic Interaction	0.08	0.09	0.00	0.08	-0.12	0.13	-0.19	0.20
Cubic	0.07	0.07	-0.02	0.07	-0.11	0.10	-0.20	0.17
Cubic Interaction	-0.04	0.14	-0.06	0.10	-0.11	0.18	0.20	0.29

* Significance level of 0.05

Next, the average residuals for six regression models considered in this study across the AZELLA binding scores for the analytic sample were plotted and visually inspected to determine the final model among them. Each of the ELL and RFEP groups in the analytic sample was divided into eleven groups (bins) with respect to the AZELLA binding score rounded to the 1st decimal place (i.e., the rounded AZELLA binding scores range from – 1 to 0 with a increment of 0.1 for the ELL group and from 0 to 1 with a increment of 0.1 for the RFEP group). The average residuals were calculated for the eleven bins for the ELL and RFEP groups. The plots of average residuals in Appendix C showed that the average residuals were generally similar and around 0 across the AZELLA binding scores across all regression models for each grade and subject combination, except for Grade 8 Reading and Mathematics. The average residual plots did not manifest a systematic pattern of positive or negative residuals for any grade and subject combination. These findings indicated that all regression models fit the merged data equally. In this case, the Linear model was picked as the final model because it was the most parsimonious model among six regression models considered in this study.

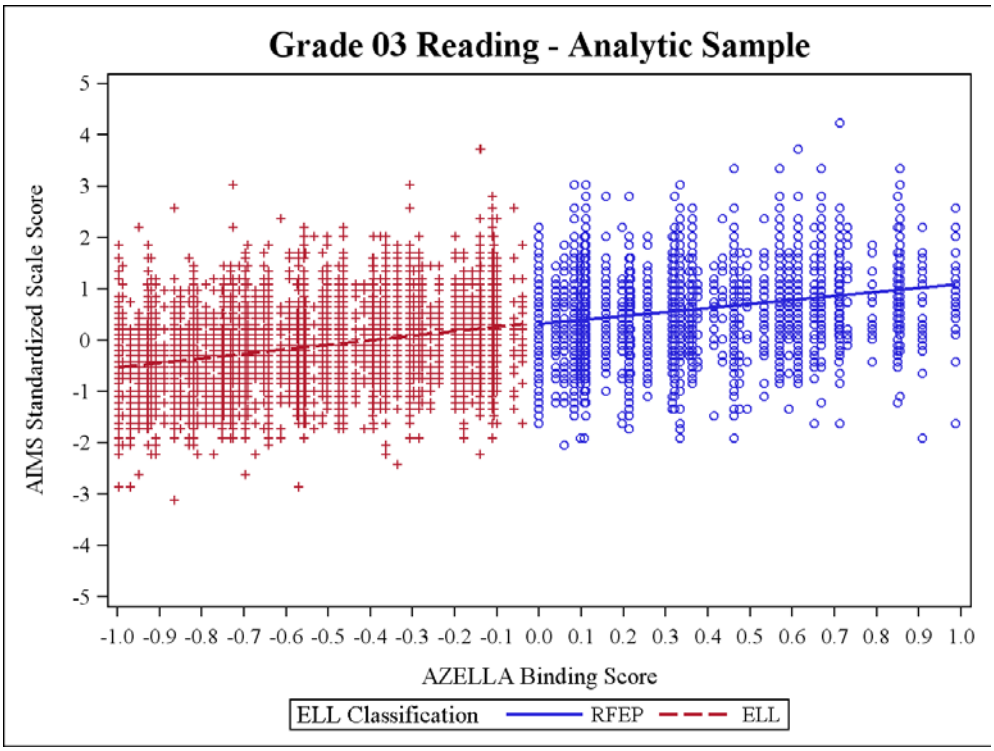
For Grade 8 Reading, the average residual plots for the Linear model was slightly away from the other regression models from the AZELLA binding score of 0 to 1. However, there was no trend that had a smaller average residual at each bin in the range of AZ binding scores between the Linear model and the other regression models. Thus, the Linear model was selected as the final model for Grade 8 Reading. Similarly, the average residual plots for the Cubic Interaction model had slightly different pattern from the other models between the AZELLA binding scores of 0 and 1 for Grade 8 Mathematics. However, the average residuals for

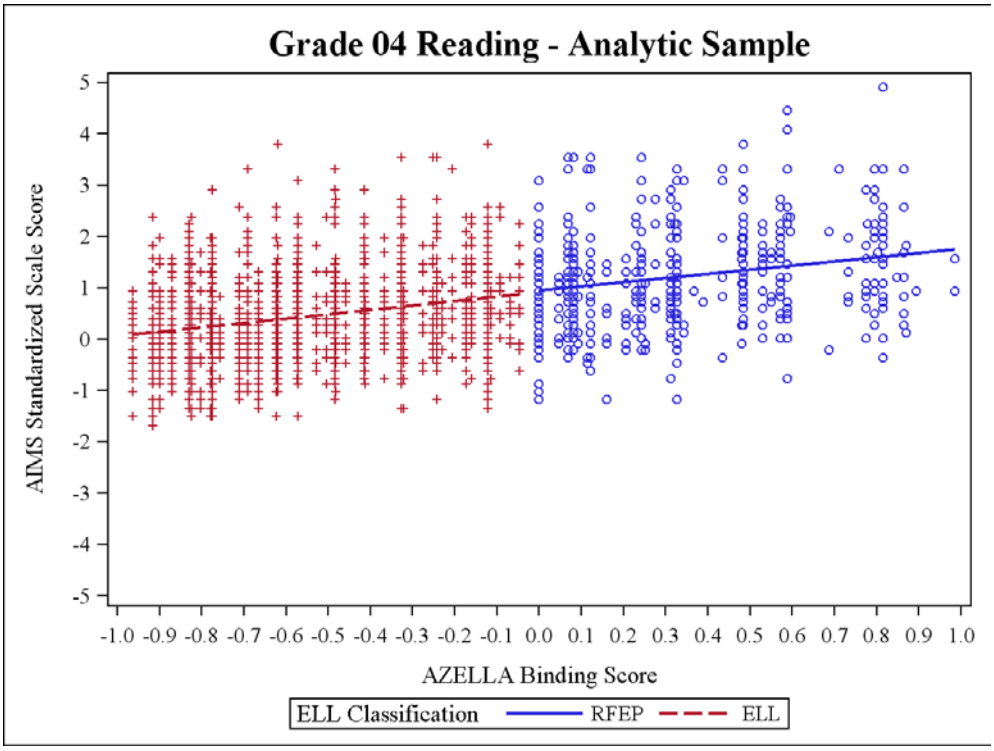
the Cubic interaction model were higher than the other models for some bins while they were lower for the other bins. Thus, it was not determined that the Cubic Interaction model was better than the others.

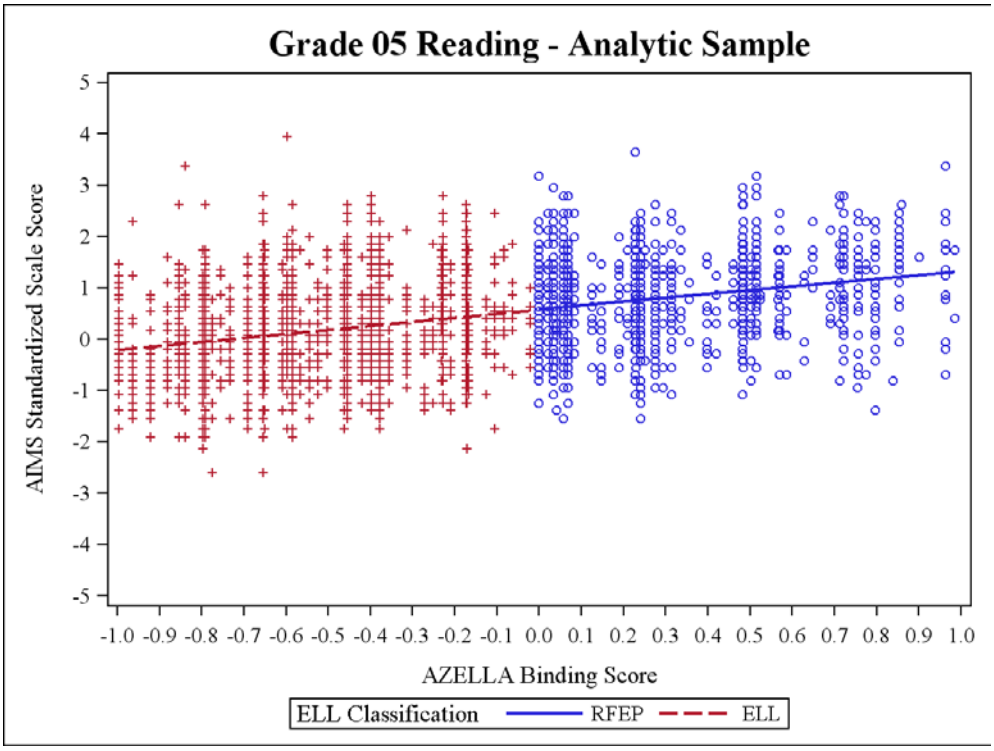
Therefore, the Linear model was determined to be the final model for all grade and subject combinations. Note that there were some average residuals that were far from 0 (e.g., average residuals close to 1); however, they were mostly due to a small sample size (N < 10) at the bins.

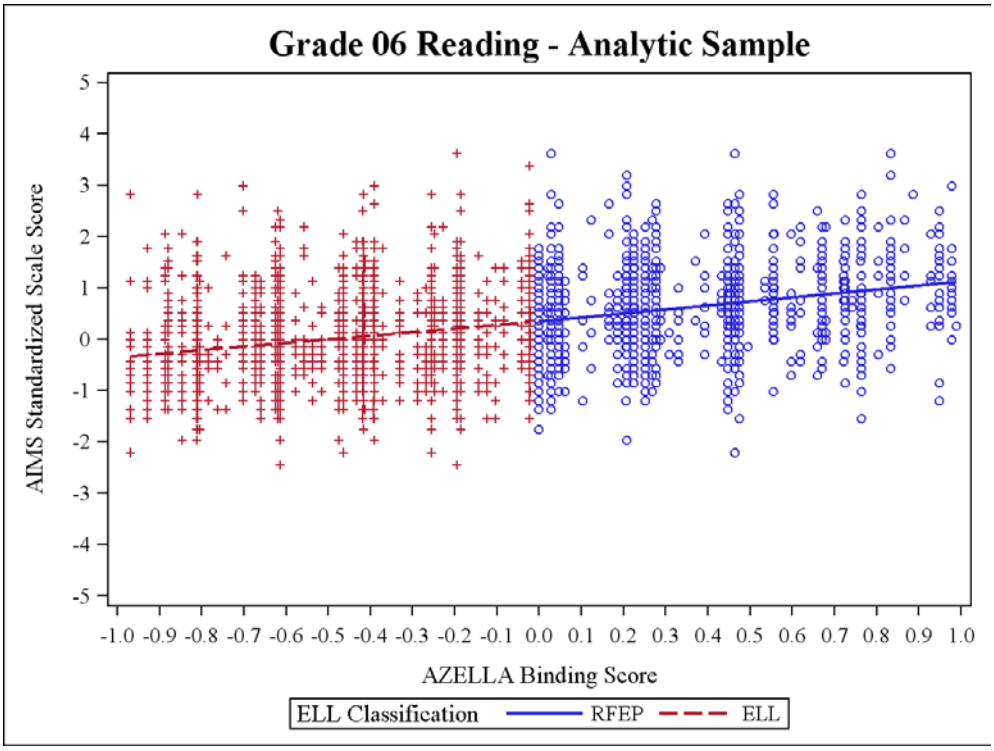
Appendix B: Scatter Plots of AIMS Scores against AZELLA Binding Scores

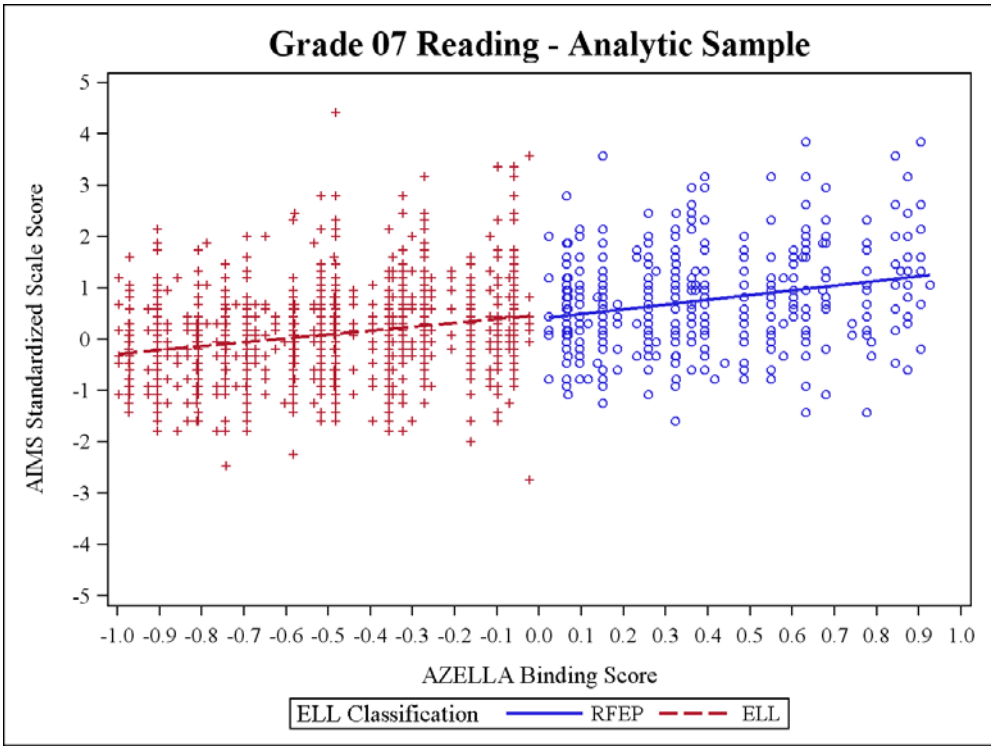
Appendix B-1: Scatter Plots of AIMS Reading Scores against AZELLA Binding Scores

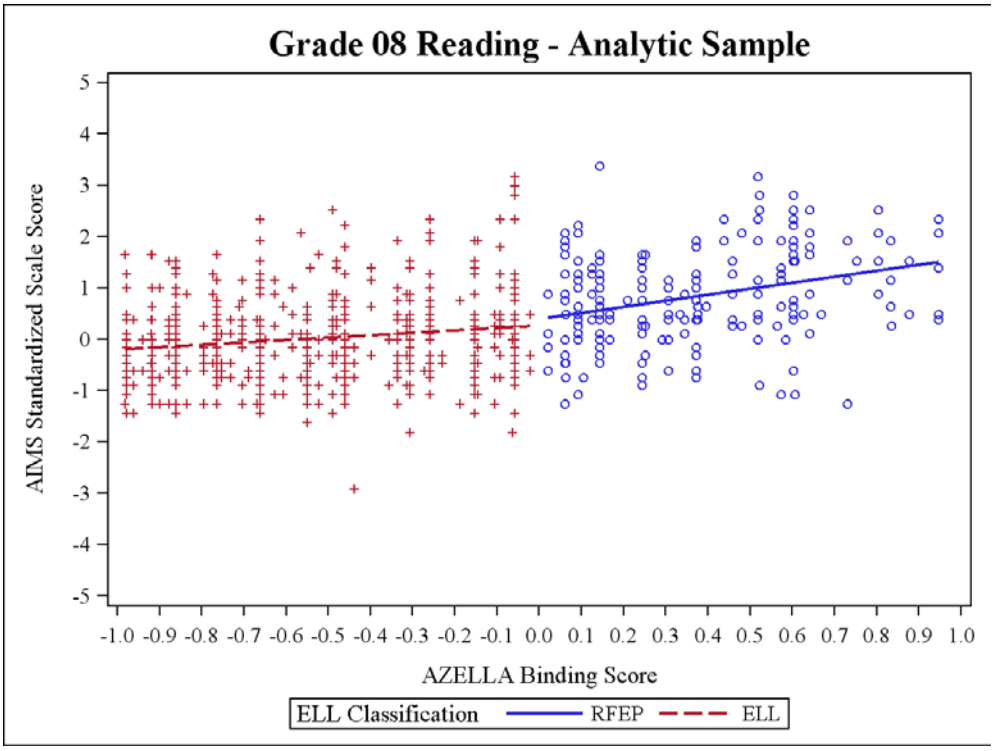


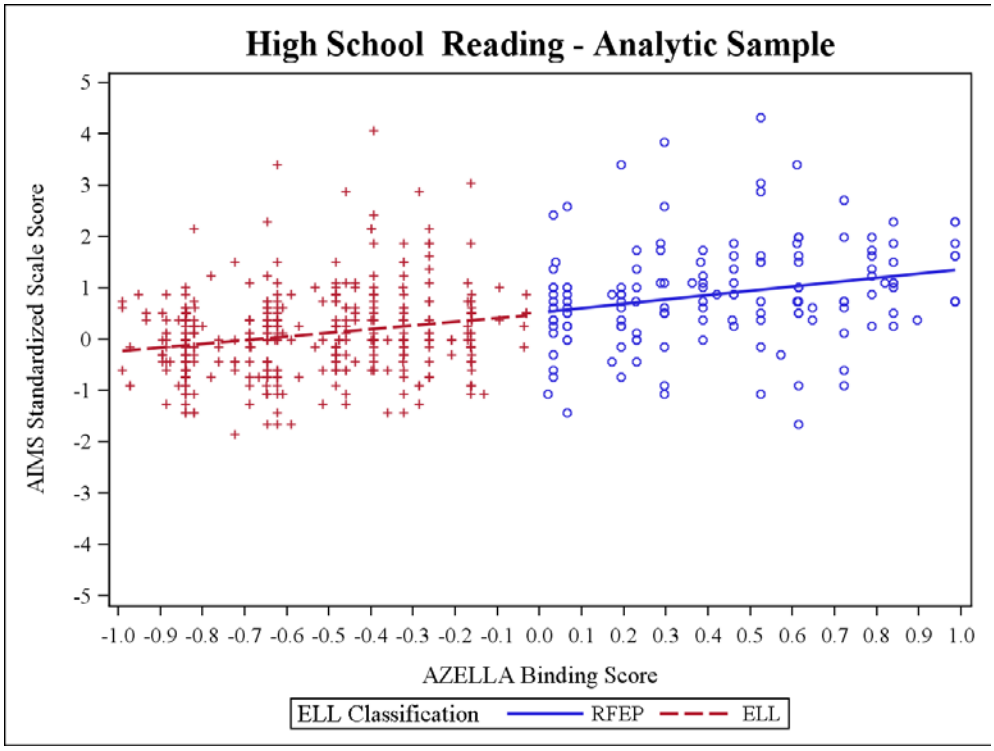



Grade 05 Reading - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL

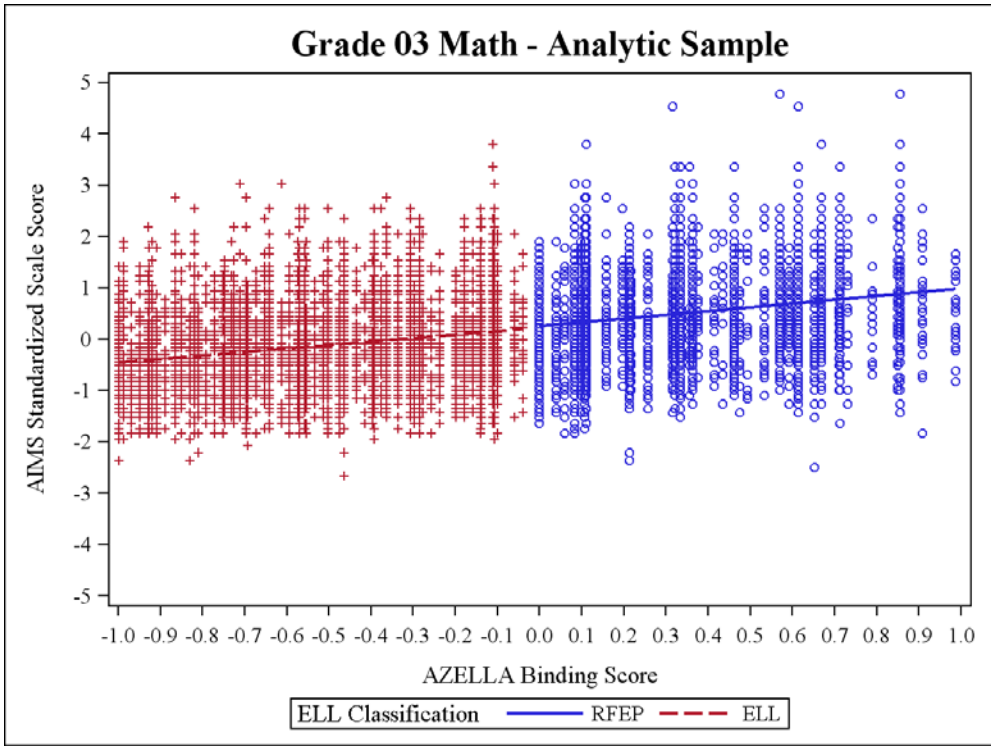

Grade 06 Reading - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL

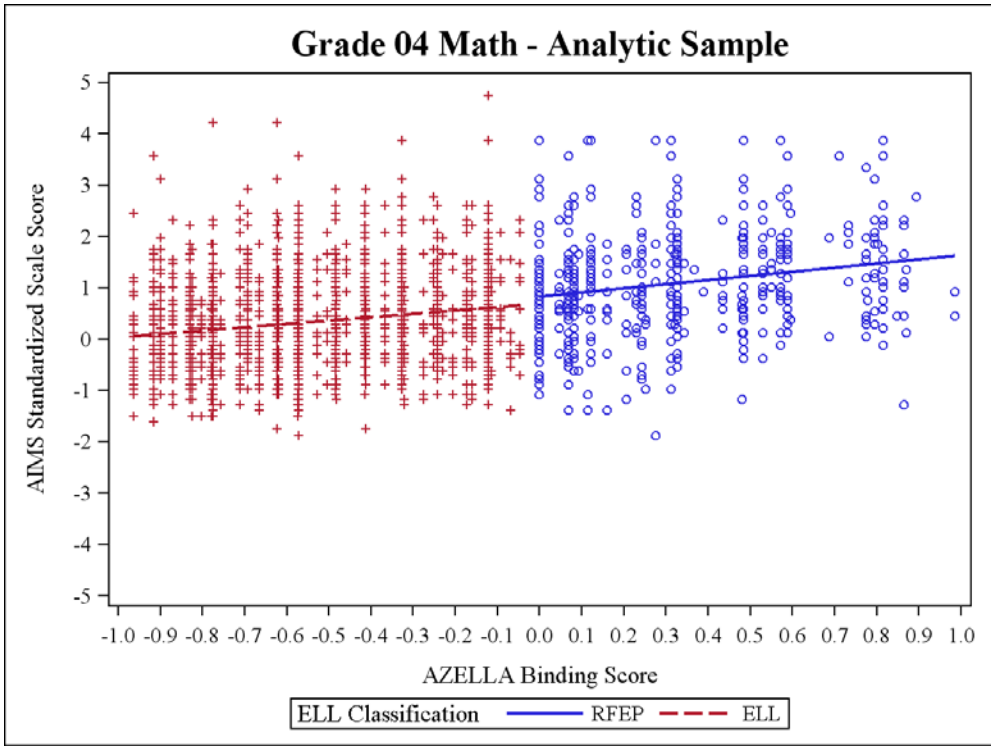

Grade 07 Reading - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL

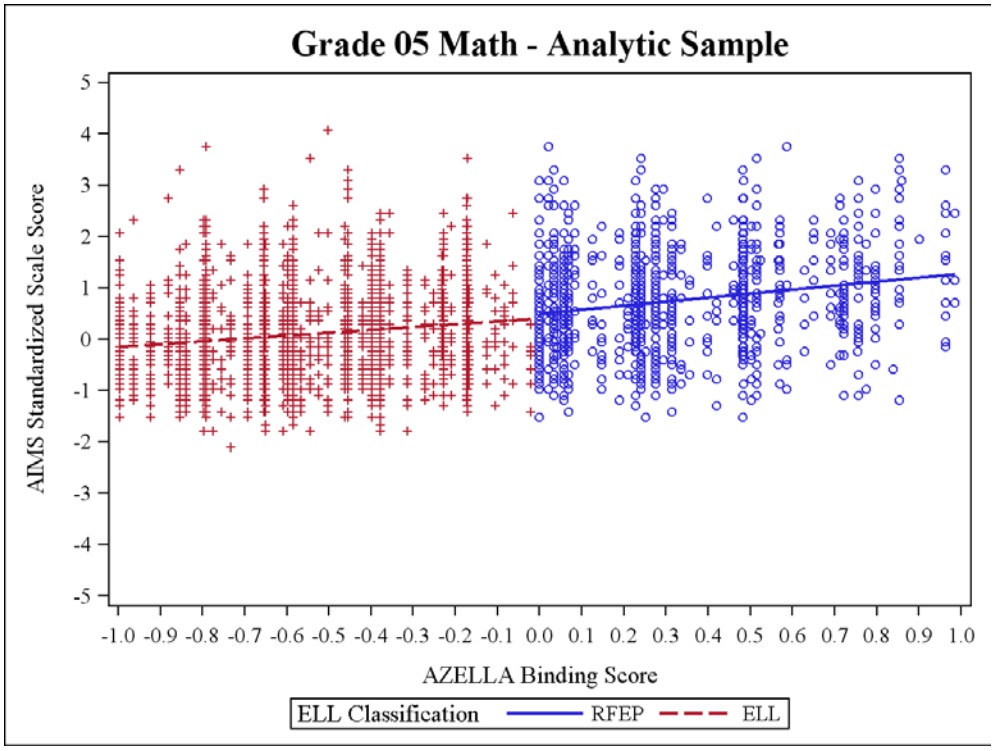

Grade 08 Reading - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL

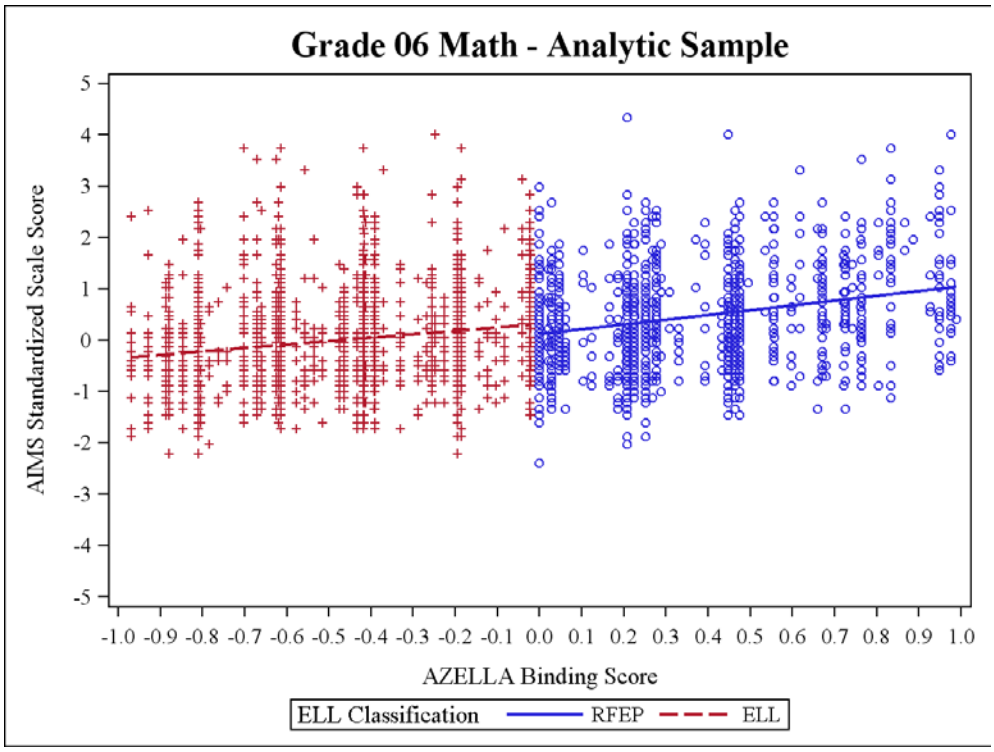

High School Reading - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL

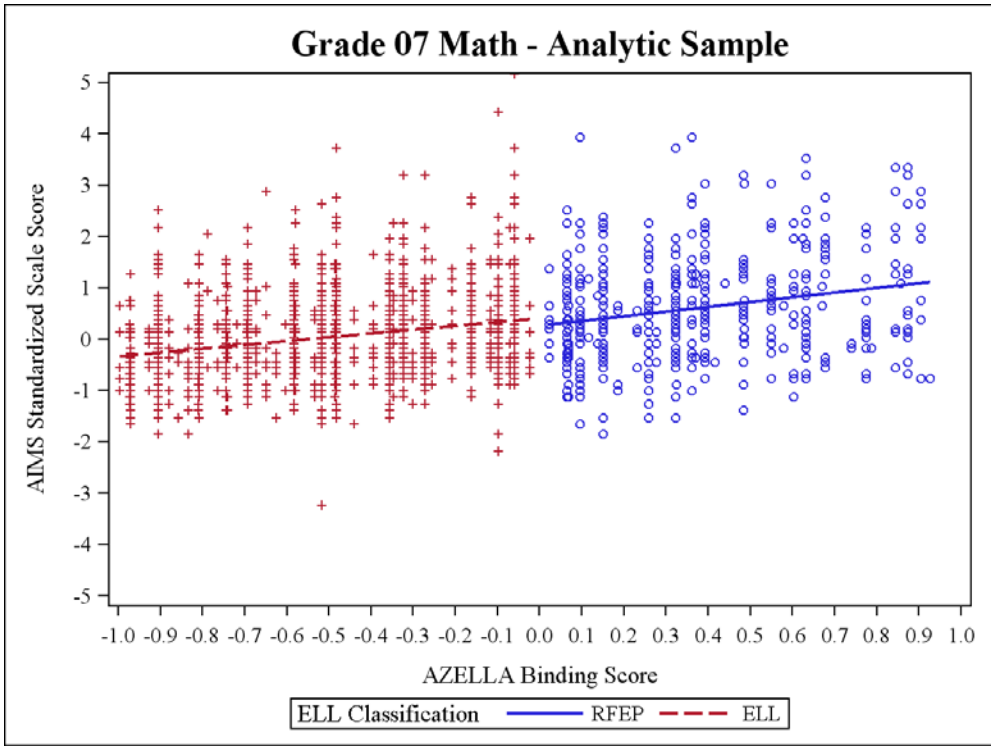
Appendix B-2: Scatter Plots of AIMS Mathematics Scores against AZELLA Binding Scores

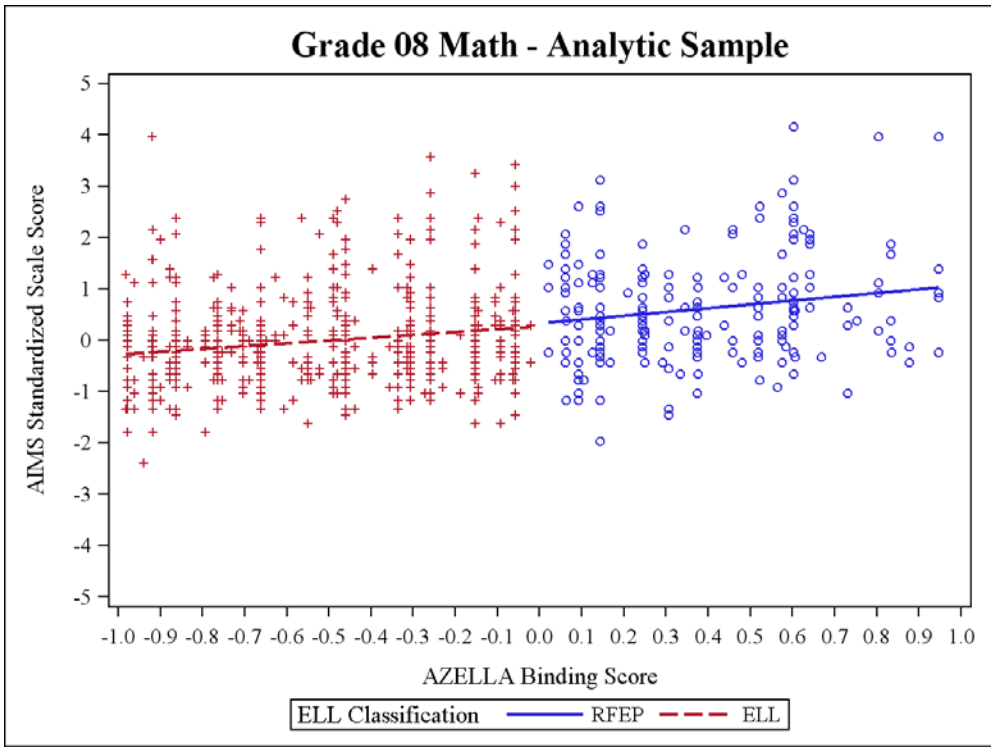


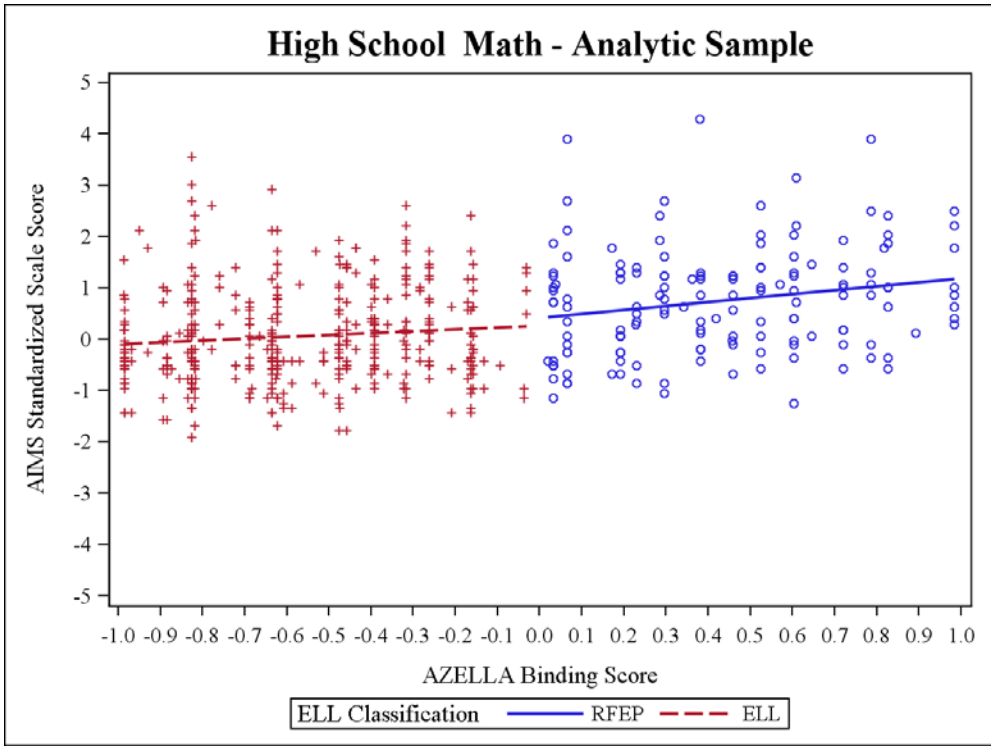



Grade 05 Math - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL

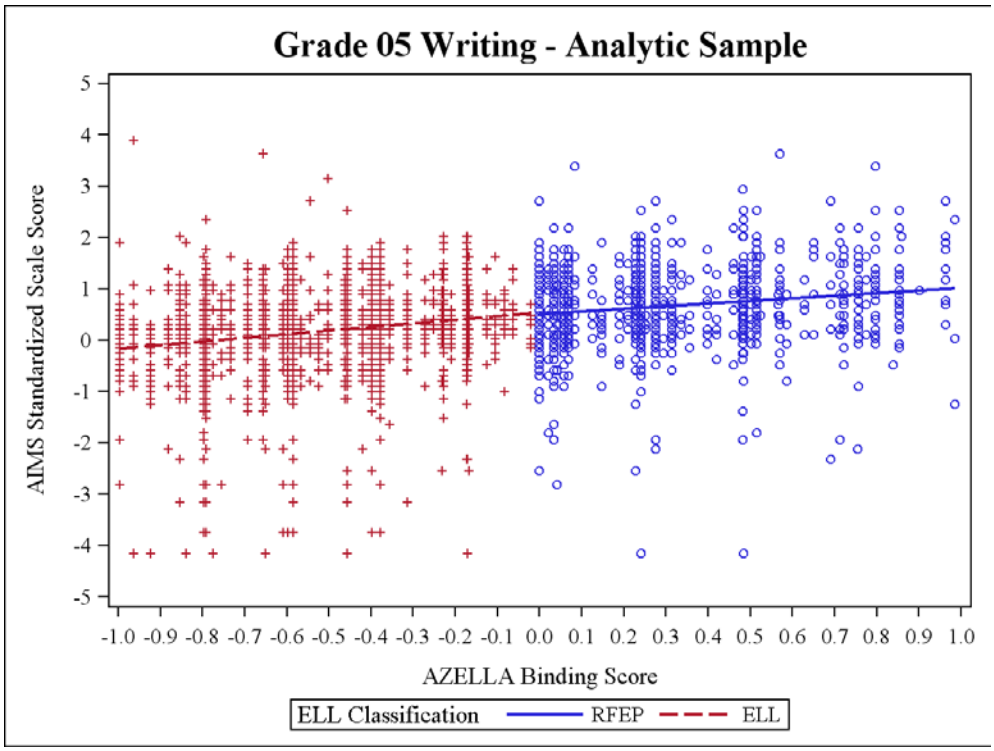

Grade 06 Math - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL

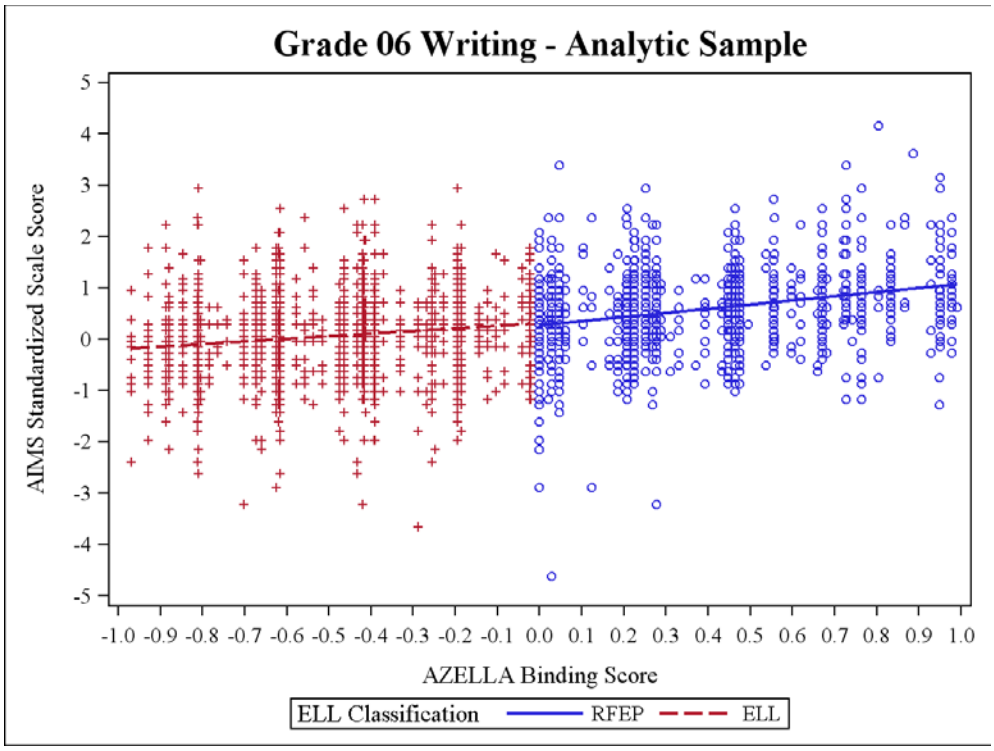

Grade 07 Math - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL

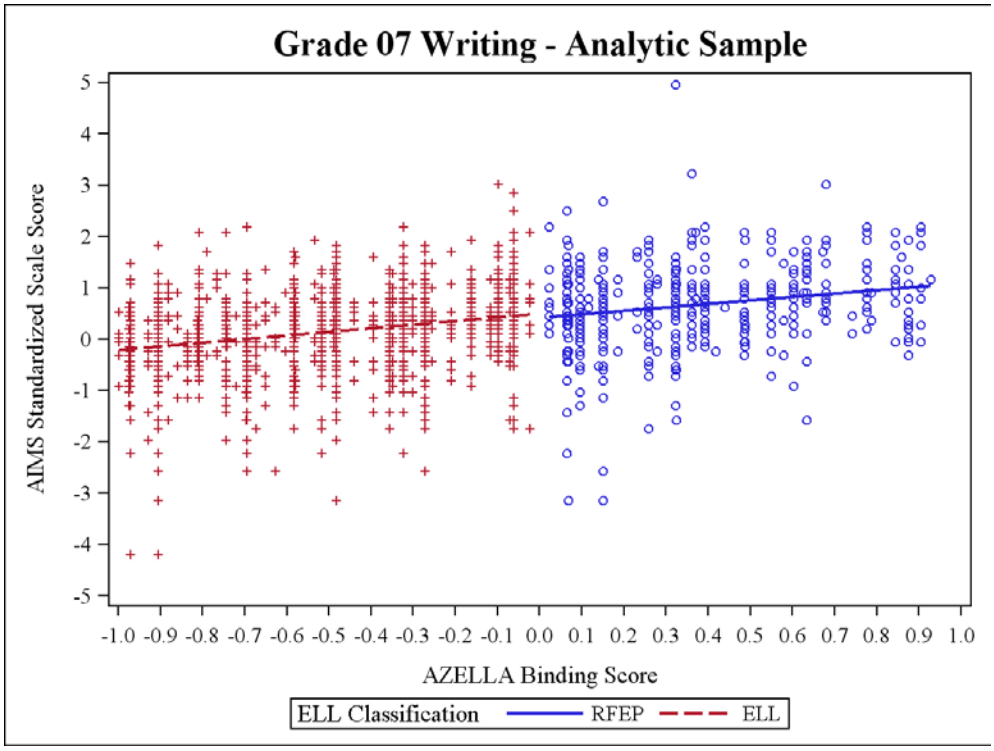

Grade 08 Math - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL

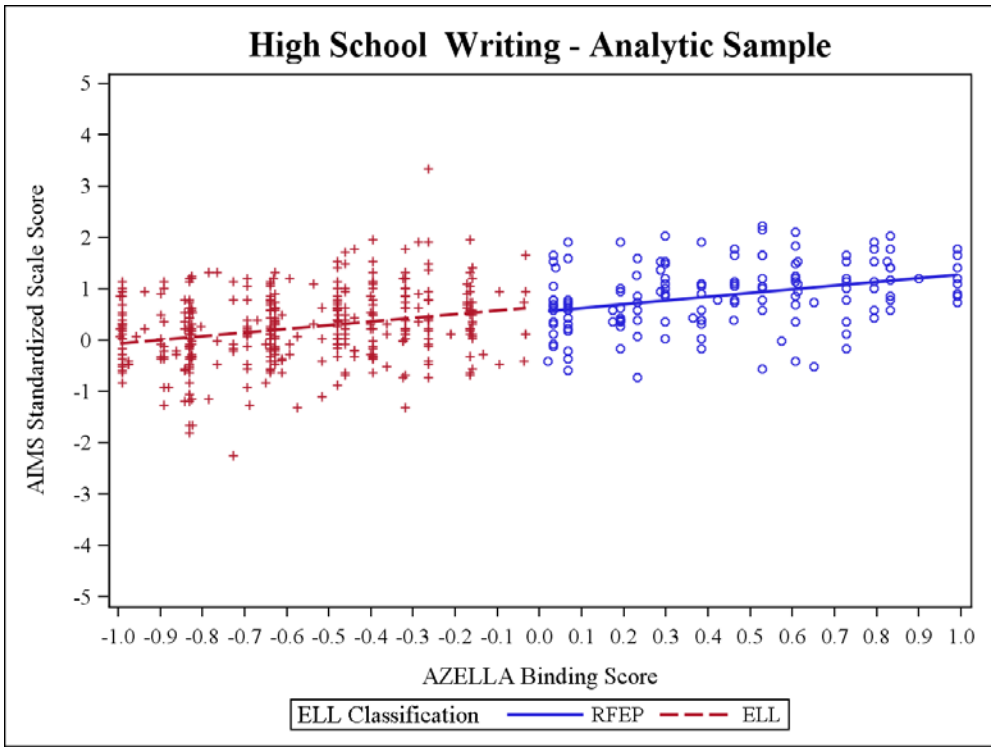

High School Math - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL

Appendix B-3: Scatter Plots of AIMS Writing Scores against AZELLA Binding Scores



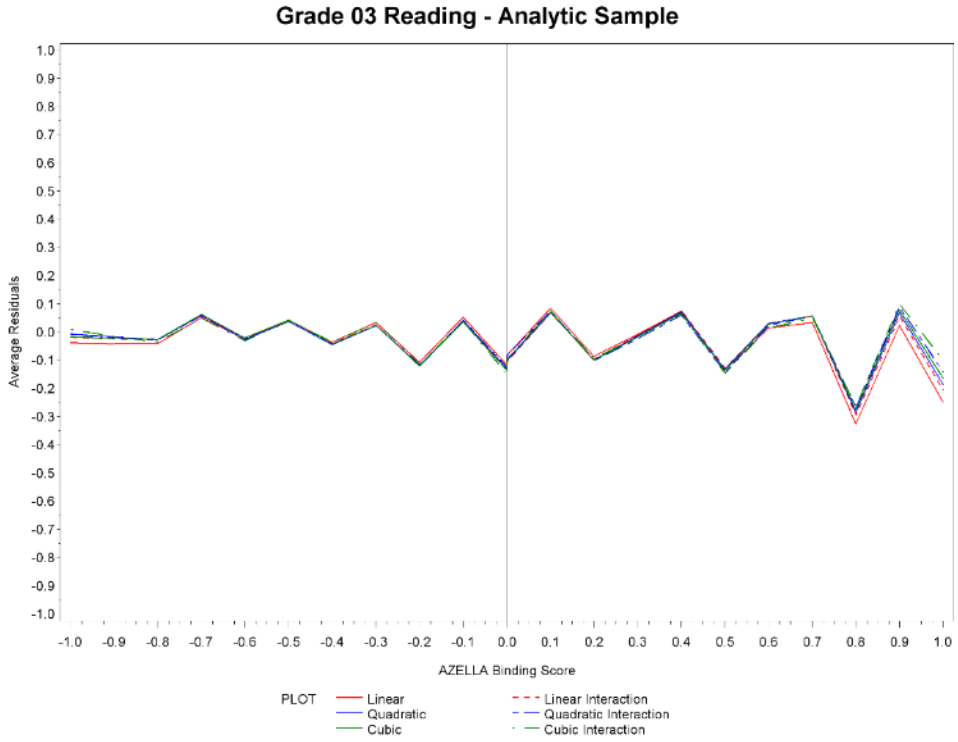


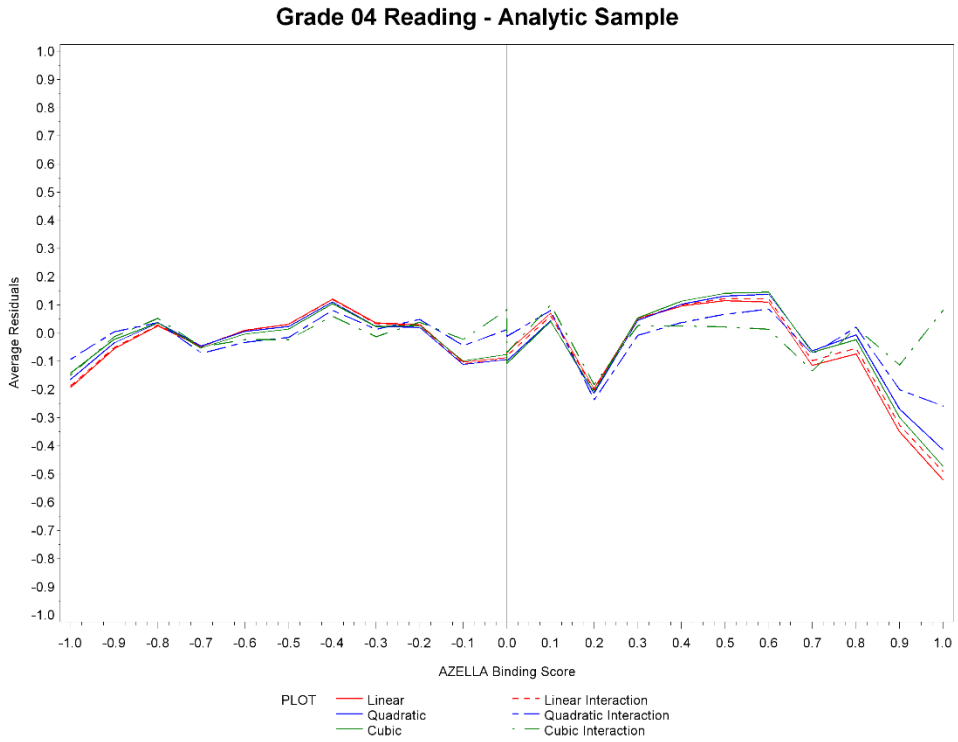

Grade 07 Writing - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL


High School Writing - Analytic Sample
AIMS Standardized Scale Score
5
4
3
2
1
0
-1
-2
-3
-4
-5
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
ELL Classification
RFEP
ELL

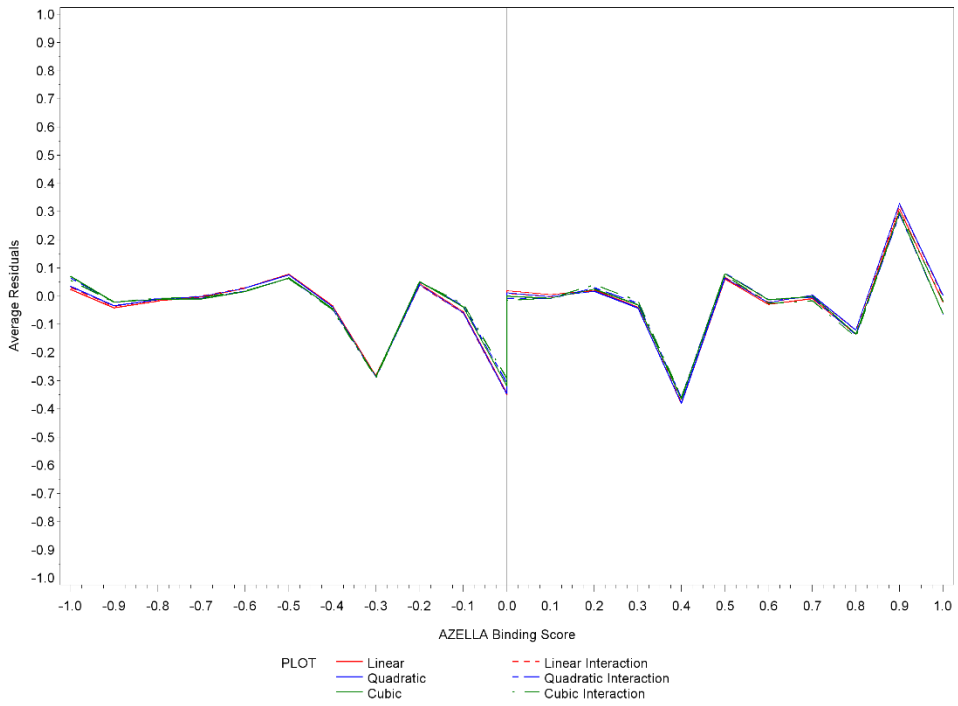
Appendix C: Average Residual Plots

Appendix C-1: Average Residual Plots of Six Regression Models for AIMS Reading

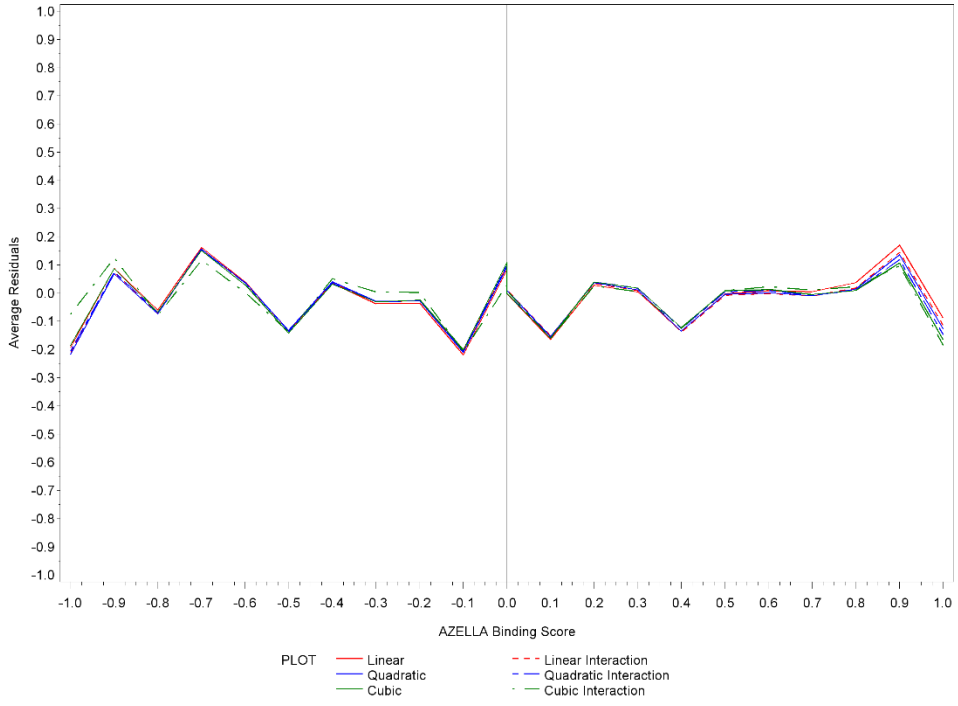




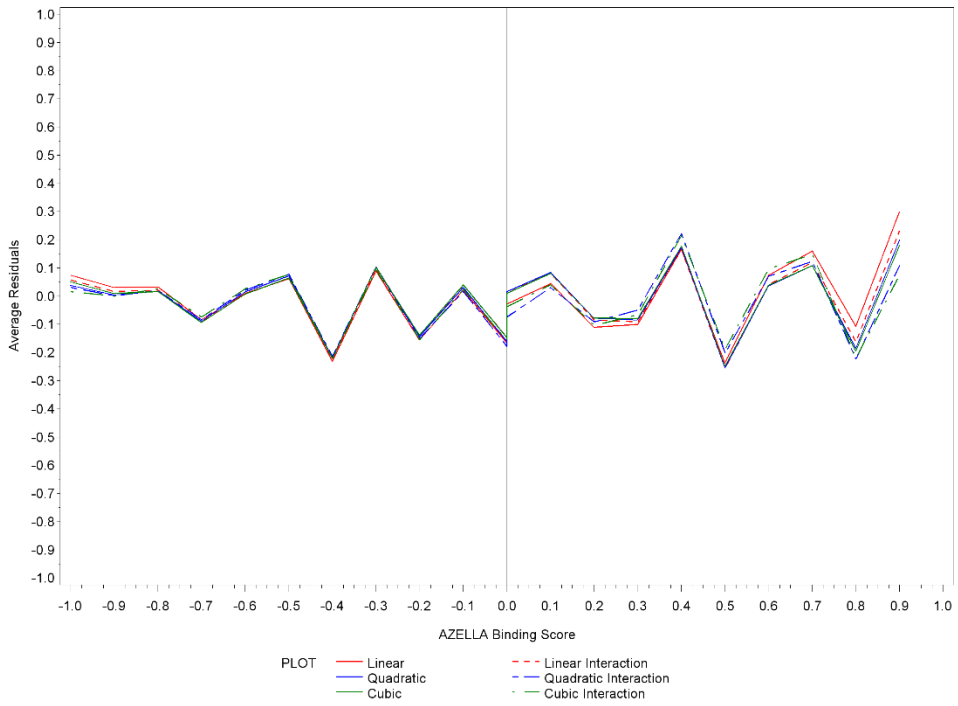
Grade 05 Reading - Analytic Sample



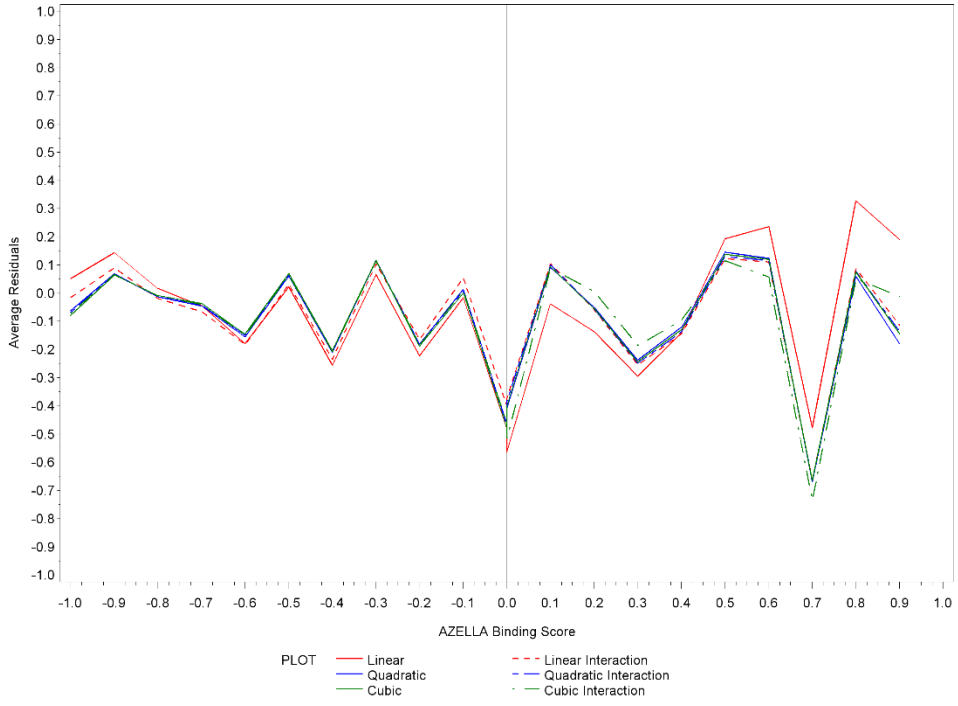
Grade 06 Reading - Analytic Sample

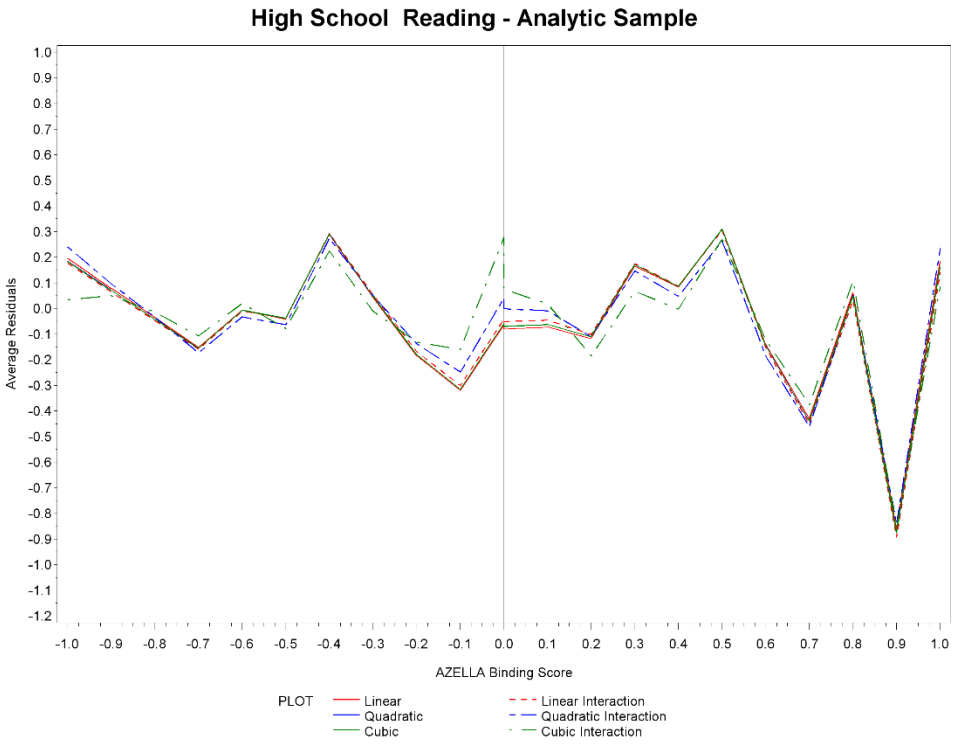


Grade 07 Reading - Analytic Sample

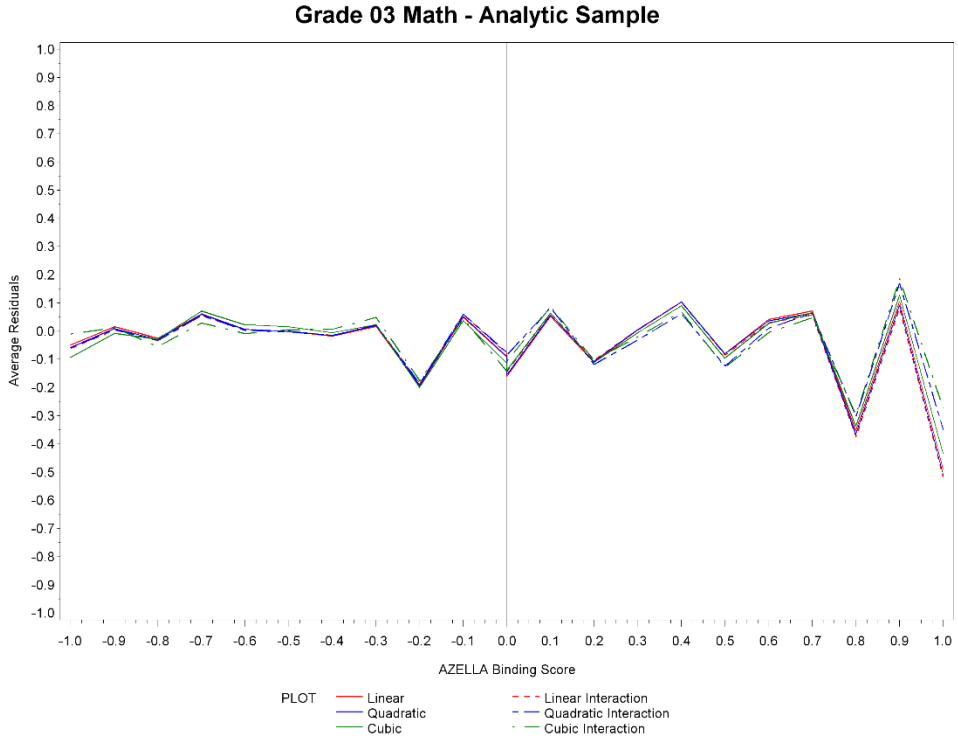


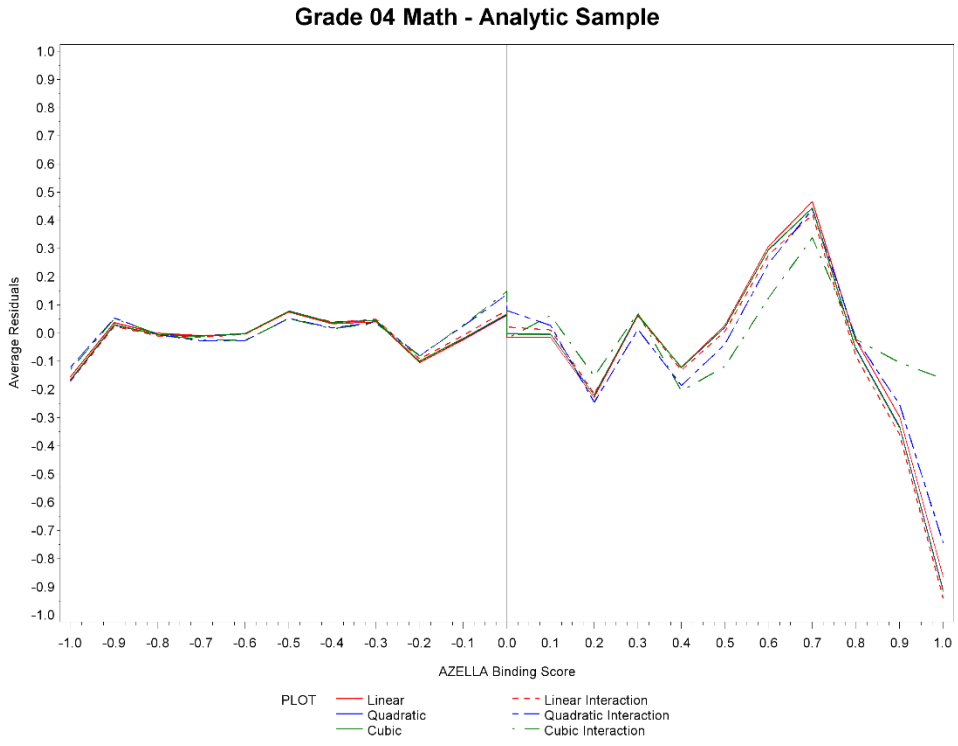
Grade 08 Reading - Analytic Sample



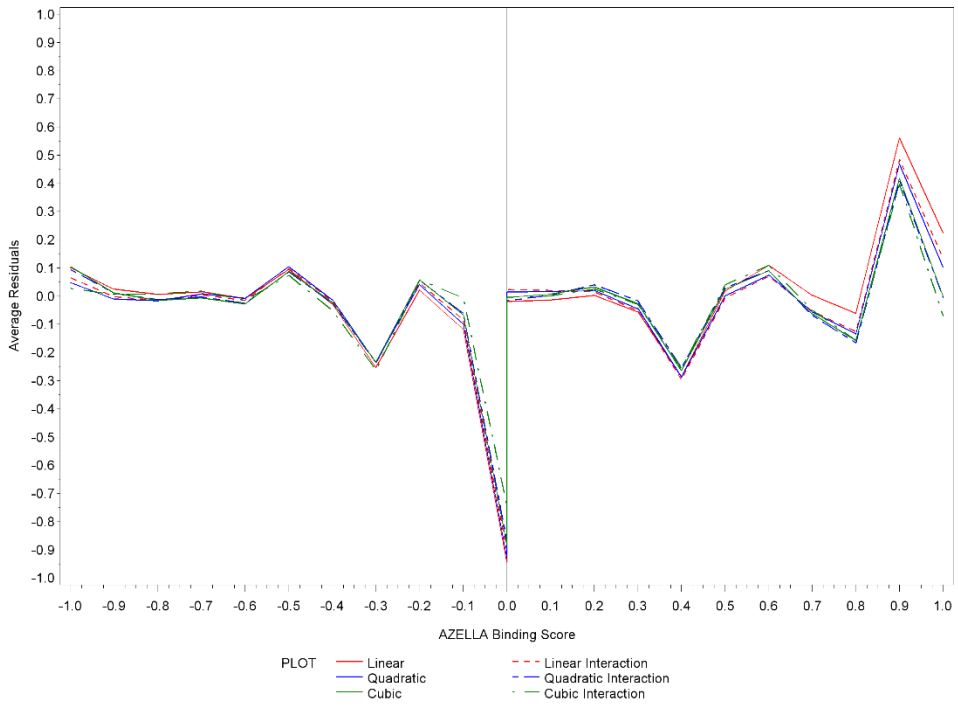

High School Reading - Analytic Sample
Average Residuals
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1.0
-1.1
-1.2
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
PLOT
Linear
Quadratic
Cubic
Linear Interaction
Quadratic Interaction
Cubic Interaction

Appendix C-2: Average Residual Plots of Six Regression Models for AIMS Mathematics

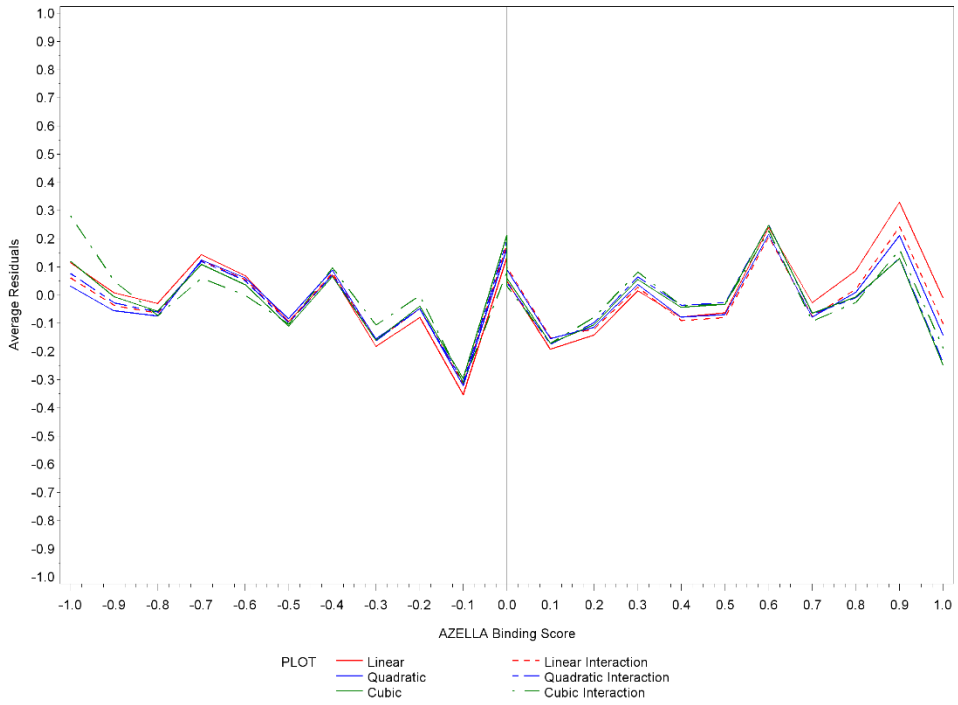




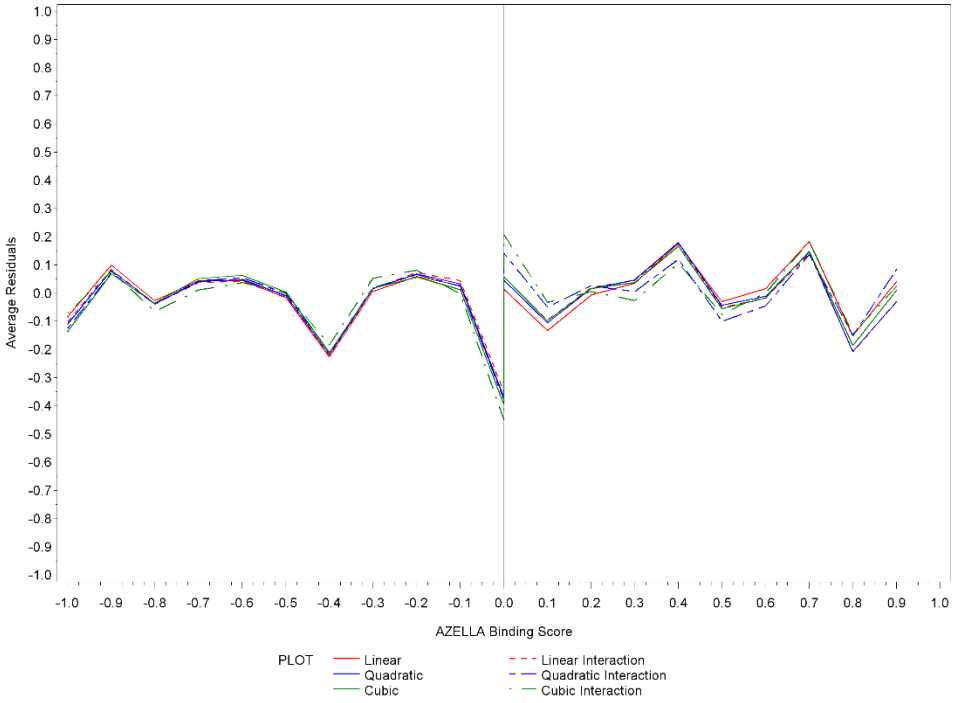
Grade 05 Math - Analytic Sample



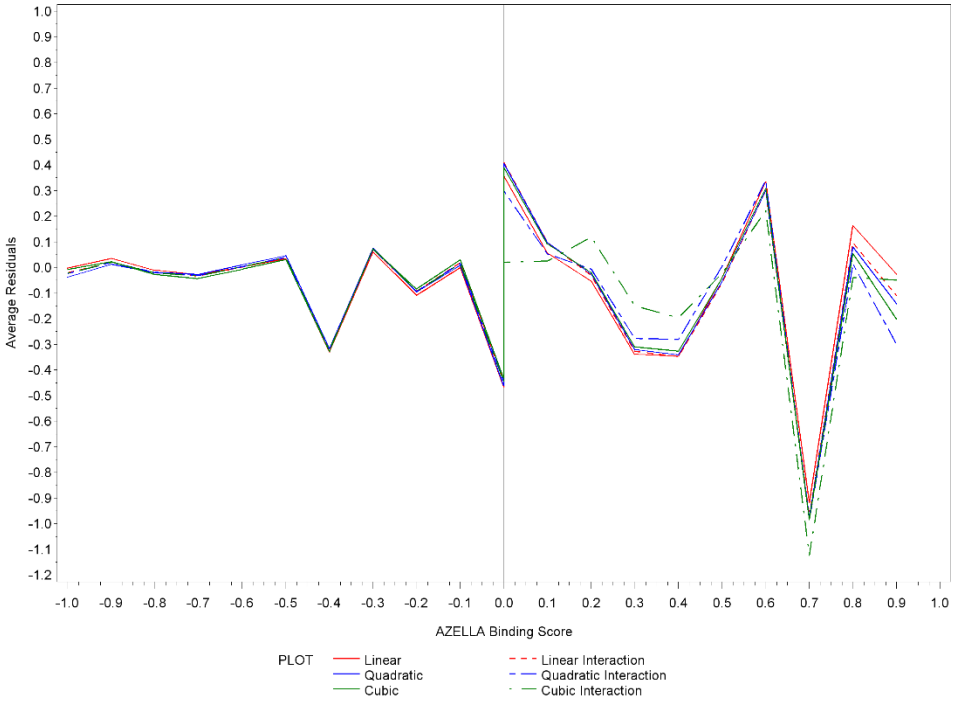
Grade 06 Math - Analytic Sample

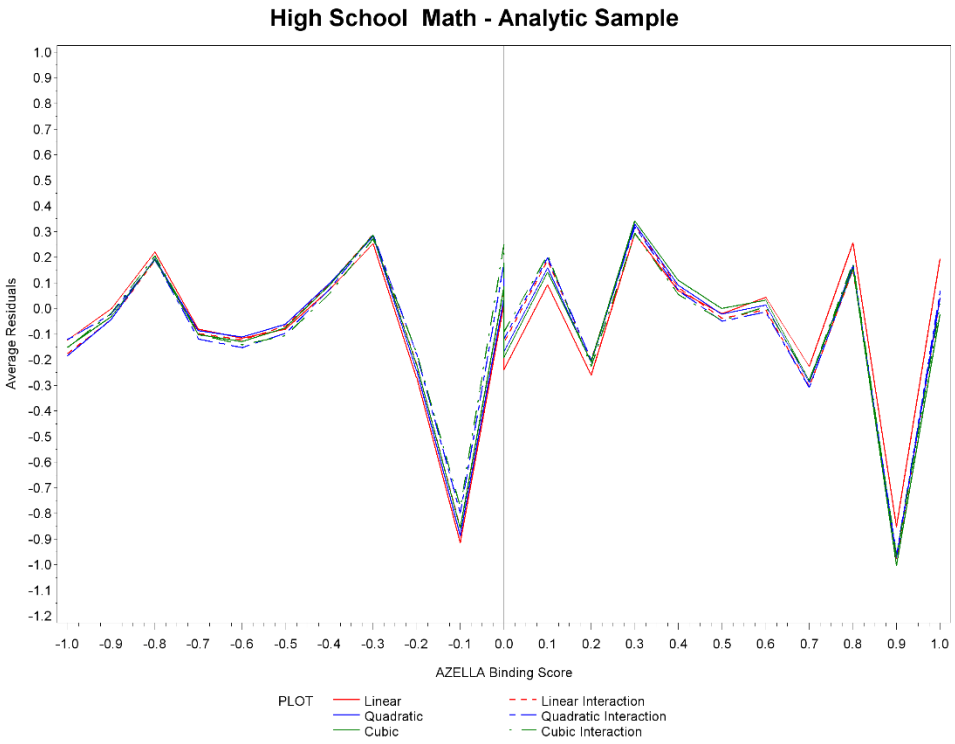


Grade 07 Math - Analytic Sample

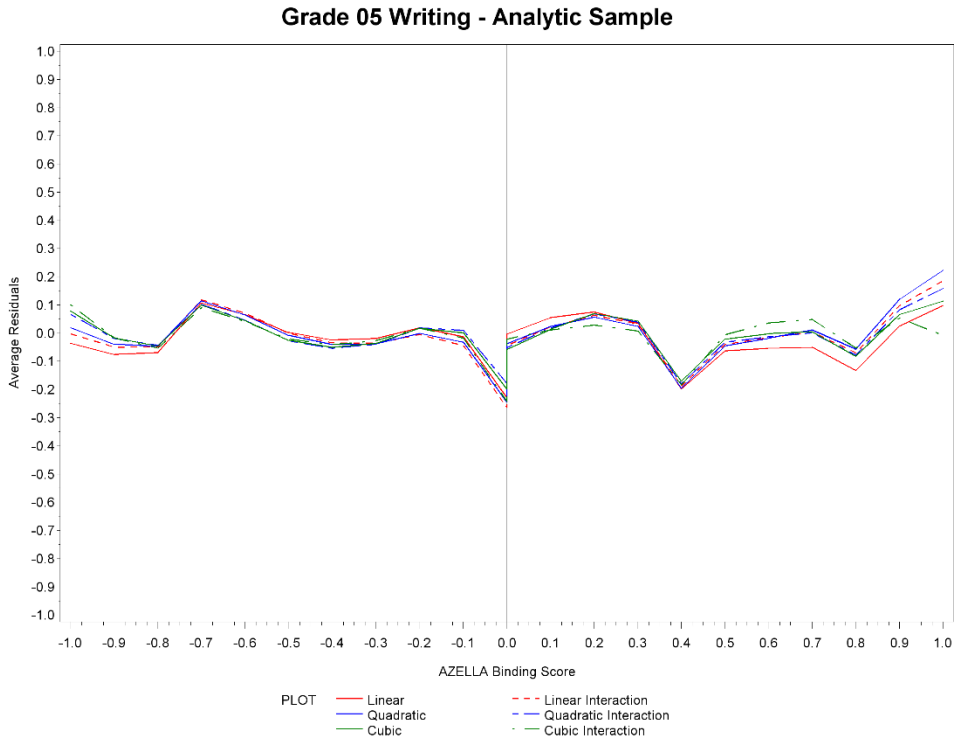


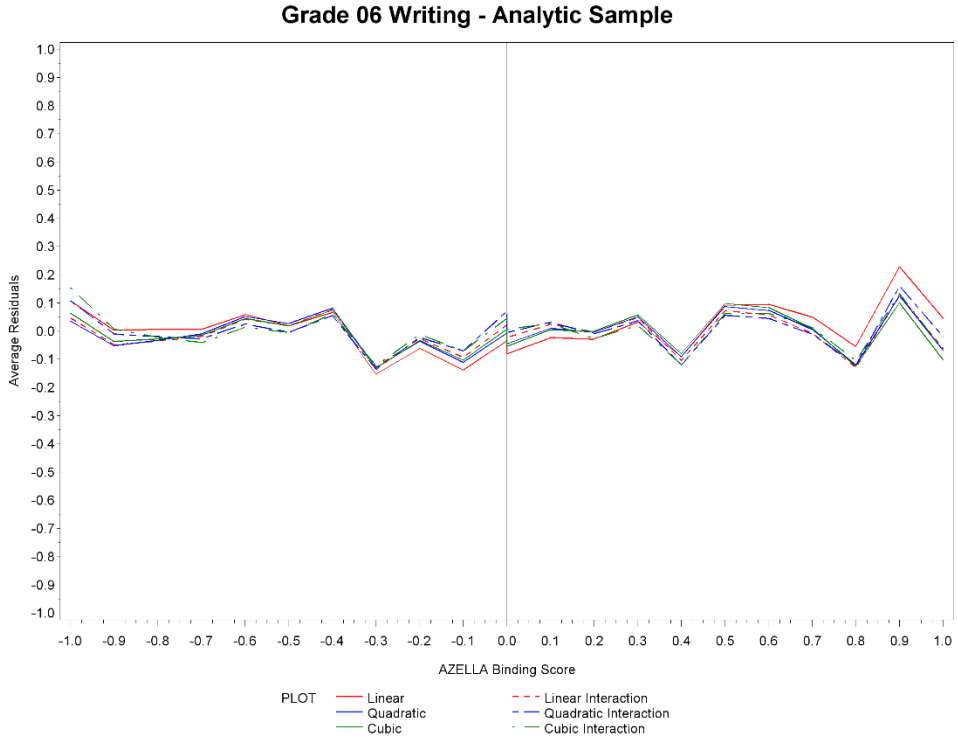
Grade 08 Math - Analytic Sample



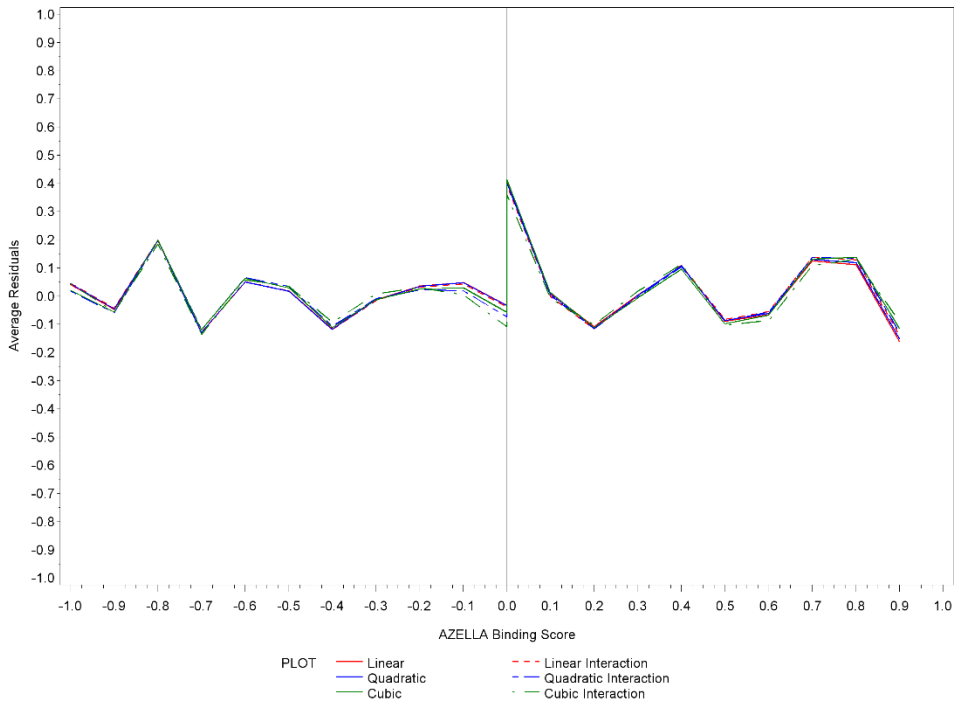

High School Math - Analytic Sample
Average Residuals
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1.0
-1.1
-1.2
-1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AZELLA Binding Score
PLOT
Linear
Quadratic
Cubic
Linear Interaction
Quadratic Interaction
Cubic Interaction

Appendix C-3: Average Residual Plots of Six Regression Models for AIMS Writing

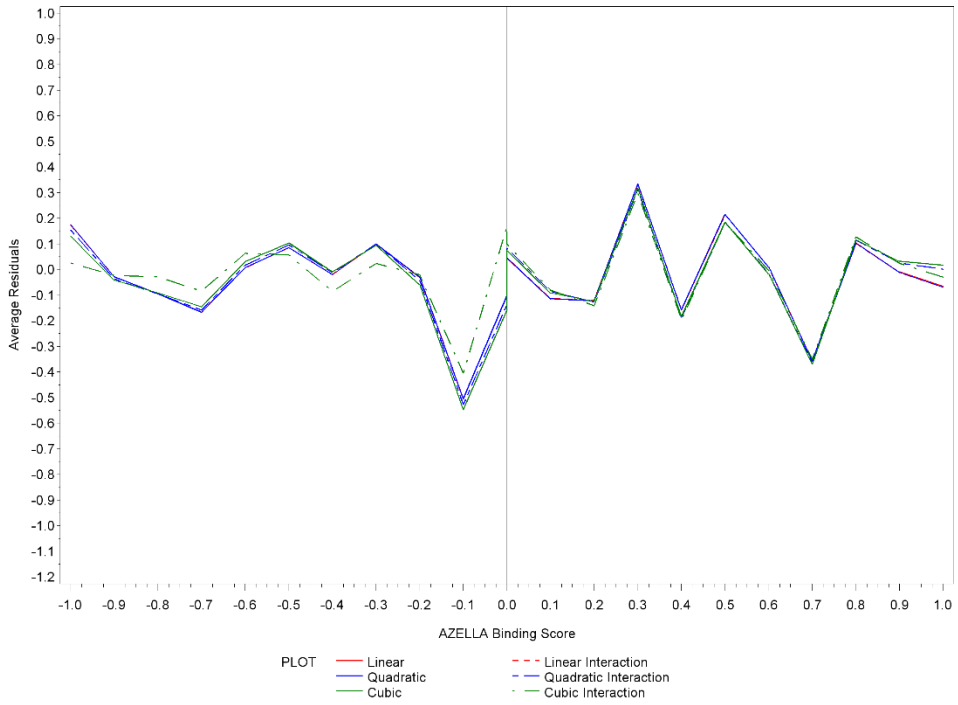




Grade 07 Writing - Analytic Sample



High School Writing - Analytic Sample



Appendix D: Sensitivity Analysis

A sensitivity analysis was conducted by adding demographic variables (i.e., gender, Hispanic, free/reduced lunch, and special education) into the final regression model as covariates for each of grade and subject combinations. If the demographic variables were not statistically significant, they were dropped from the model. The parameter estimates of ELL classification (δ) between the final model with and without the demographic variables are presented in Tables C.1 through C.3. The sensitivity analysis did not lead to different results between the final model with and without the demographic variables for any grade and subject combination, except for Grade 8 Reading, in which the parameter estimate became non-significant after the demographic variables were included in the analysis. Note that the sample size was slightly decreased for the sensitivity analysis due to missing demographic variables for some students in the analytic sample.

Table C.1: Parameter Estimate of ELL Classification on AIMS Reading in Final Model and Final Model with Demographic Variables

			Final Model			Final Model with Demographics		
Grade	Final Model	+Demographics Included	*N*	δ	*SE*	*N*	δ	*SE*
3	Linear	G, H, F, S	8219	-0.05	0.03	8207	-0.04	0.03
4	Linear	F, S	1983	0.03	0.08	1981	0.02	0.08
5	Linear	H, F, S	3148	0.01	0.06	3140	0.01	0.06
6	Linear	G, H, F, S	2848	-0.01	0.06	2841	0.00	0.06
7	Linear	G, H, F, S	1635	-0.05	0.08	1631	-0.04	0.08
8	Linear	G, S	741	*0.25	0.13	739	0.23	0.12
High School	Linear	S	493	0.04	0.17	472	0.05	0.17

+G: Gender, H: Hispanic, F: Free/reduced Lunch, S: Special Education
* Significance level of 0.05

Table C.2: Parameter Estimate of ELL Classification on AIMS Mathematics in Final Model and Final Model with Demographic Variables

			Final Model			Final Model with Demographics		
Grade	Final Model	+Demographics Included	*N*	δ	*SE*	*N*	δ	*SE*
3	Linear	G, H, S	8222	0.04	0.03	8210	0.04	0.03
4	Linear	G, H, F, S	1983	0.16	0.09	1981	0.14	0.09
5	Linear	G, H, F, S	3148	0.11	0.06	3139	0.11	0.06
6	Linear	H, F, S	2847	*-0.19	0.07	2841	*-0.18	0.06
7	Linear	H, F, S	1634	-0.11	0.09	1630	-0.09	0.09
8	Linear	H, S	740	0.10	0.14	738	0.10	0.14
High School	Linear	H, S	526	0.19	0.18	505	0.13	0.18

+G: Gender, H: Hispanic, F: Free/reduced Lunch, S: Special Education
* Significance level of 0.05

Table C.3: Parameter Estimate of ELL Classification on AIMS Writing in Final Model and Final Model with Demographic Variables

			Final Model			Final Model with Demographics		
Grade	Final Model	+Demographics Included	*N*	δ	*SE*	*N*	δ	*SE*
5	Linear	G, H, F, S	3146	-0.03	0.06	3137	-0.01	0.05
6	Linear	G, H, F, S	2838	-0.04	0.06	2831	-0.03	0.05
7	Linear	G, H, F, S	1630	-0.08	0.08	1627	-0.09	0.07
High School	Linear	S	522	-0.10	0.12	501	-0.12	0.12

+G: Gender, H: Hispanic, F: Free/reduced Lunch, S: Special Education

* Significance level of 0.05