Listening. Learning. Leading.®

Research Report
ETS RR-12-07

Evaluating Academic Progress Without a Vertical Scale

Wendy M. Yen

Venessa F. Lall

Lora Monfils

April 2012

Evaluating Academic Progress Without a Vertical Scale

Wendy M. Yen, Venessa F. Lall, and Lora Monfils
ETS, Princeton, New Jersey

April 2012

Technical Review Editor: Daniel Eignor

Technical Reviewers: Henry Braun and Matthias von Davier

Abstract

Alternatives to vertical scales are compared for measuring longitudinal academic growth and for producing school-level growth measures. The alternatives examined were empirical cross-grade regression, ordinary least squares and logistic regression, and multilevel models. The student data used for the comparisons were Arabic Grades 4 to 10 in Qatar, and results were examined in the scale score and performance level metrics. It is found that vertical scales and cross-grade regressions can show different results at the individual student level, but at the school level, the different measures of growth were strongly correlated, particularly in the scale score metric. Differences between the methods appear more likely in the performance level metric than the scale score metric and for grade pairs with more extreme performance.

Key words: growth, longitudinal analyses, vertical scaling

Acknowledgments

We thank Robert Smith, ETS, for drafting an introduction to this paper; Lu Ying, ETS, for providing SAS code for the pooled adjacent violators algorithm; and Lixiong Gu, ETS, for his assistance with the statistical analyses associated with the paper.

Table of Contents

Page

Method 2

Data Source 2

Performance Levels 3

Academic Growth Measures 9

Results 14

Empirical Regression 14

Ordinary Least Squares and Logistic Regression 14

Multilevel Models 20

Comparisons of School Level Results Across Methods 26

Summary and Discussion 45

Applicability of Each Method 46

Summary 48

References 53

Notes 55

List of Tables

Page

Table 1. Scale Score Summary Statistics for Merged Records—Arabic 4

Table 2. Scale Score Summary Statistics for Merged Records—English 5

Table 3. Sample Size Summary for 2006 Schools by Grade 6

Table 4. Approaches Standards Cut Scores for Arabic and English.. 7

Table 5. Performance Level Summary Statistics for Merged Records—Arabic 8

Table 6. Performance Level Summary Statistics for Merged Records—English...................... 8

Table 7. Cross-Grade Correlations of Student-Level Scores for Arabic and English 9

Table 8. Results of Fitting Ordinary Least Squares Regression Models to Project 2006 Scale Score From 2005 Scale Scores 17

Table 9. Results of Fitting Logistic Regression Models to Project 2006 Performance Level From 2005 Scale Scores 20

Table 10. Multilevel Linear Model Results for Scale Scores 22

Table 11. Multilevel Logistic Model Results for Percentages Approaches Standards 27

Table 12. Means, Standard Deviations, and Correlations Between School Mean Growth Measures Based on Scale Score Projections.. 39

Table 13. Means, Standard Deviations, and Correlations Between School Mean Growth Measures Based on Performance Level Projections 40

Table 14. Correlations Between Growth Based on Scale Scores Projections and Growth Based on Performance Level Projections Within the Same Projection Method 42

List of Figures

Page

Figure 1. Empirical regression: 2006 projected Arabic scale scores versus 2005 Arabic scale scores. 15

Figure 2. Empirical regression: 2006 projected English scale scores versus 2005 English scale scores. 15

Figure 3. Empirical regression: 2006 projected Arabic percentage Approaches Standards versus 2005 Arabic scale scores. 16

Figure 4. Empirical regression: 2006 projected English percentage Approaches Standards versus 2005 English scale scores. 16

Figure 5. Ordinary least square regression: 2006 projected Arabic scale scores versus 2005 Arabic scale scores. 18

Figure 6. Ordinary least square regression: 2006 projected English scale scores versus 2005 English scale scores 18

Figure 7. Logistic regression: 2006 projected Arabic percentage proficient versus 2005 Arabic scale scores. 21

Figure 8. Logistic regression: 2006 projected English percentage proficient versus 2005 English scale scores 21

Figure 9. Multilevel model: 2006 projected Arabic scale scores versus 2005 Arabic scale scores. 25

Figure 10. Multilevel model: 2006 projected English scale scores versus 2005 English scale scores. 25

Figure 11. Multilevel: 2006 projected Arabic percentage Approaches Standards versus 2005 Arabic scale scores. 28

Figure 12. Multilevel: 2006 projected English percentage Approaches Standards versus 2005 English scale scores 28

Figure 13. Grade 6 Arabic 2006 projected scale scores by method. 30

Figure 14. Grade 8 English 2006 projected scale scores by method. 30

Figure 15. Grade 6 Arabic 2006 projected percentage Approaches Standards by method. 31

Figure 16. Grade 8 English 2006 projected percentage Approaches Standards by method. 31

Figure 17. Observed versus predicted school means: Arabic Grade 6 scale scores. 32

Figure 18. Observed versus predicted school means: Arabic Grade 8 scale scores. 33

Figure 19. Observed versus predicted school means: English Grade 6 scale scores.................... 33

Figure 20. Observed versus predicted school means: English Grade 8 scale scores.................... 34

Figure 21. Observed vs. predicted school percentage Approaches Standards: Arabic Grade 6. 34

Figure 22. Observed versus predicted school percentage Approaches Standards: Arabic Grade 8. 35

Figure 23. Observed versus predicted school percentage Approaches Standards: English Grade 6.. 35

Figure 24. Observed versus predicted school percentage Approaches Standards: English Grade 8.. 36

Figure 25. Bivariate plots of school mean difference between Arabic Grade 6 observed 2006 scale score and projected 2006 scale score... 43

Figure 26. Bivariate plots of school mean difference between English Grade 8 observed 2006 scale score and projected 2006 scale score... 43

Figure 27. Bivariate plots of school mean difference between Arabic Grade 6 observed 2006 percentage Approaches Standards and projected 2006 percentage Approaches Standards.. 44

Figure 28. Bivariate plots of school mean difference between English Grade 8 observed 2006 percentage Approaches Standards and projected 2006 percentage Approaches Standards.. 44

In many accountability systems, local educational authorities are required to measure student achievement with outcome measures that are tied to content and performance standards. However, beyond accountability, educators have great interest in measuring student growth over time for evaluation and program improvement purposes. In order to minimize the time students spend in testing, the same set of assessments is often asked to serve both purposes. Several models can measure academic progress. Depending on features of the accountability system and the included assessments, growth may be assessed at the school level through cross-sectional analysis of successive cohorts of students or by longitudinal analysis of the same cohort at the group or individual level.

In evaluating the effectiveness of educational programs, there are limitations to cross-sectional comparisons (e.g., comparisons of the percent of proficient students from last year's Grade 4 and this year's Grade 4). Cross-sectional comparisons are, in effect, comparing nonequivalent groups, without controlling for prior achievement or other intake variables. In sum, these methods are limited because they do not account for individual student progress but rather focus on discrete measures of group performance using nonequivalent groups (Doran & Izumi, 2004). Therefore, to evaluate the contributions of school programs to student progress, longitudinal analyses that track individual student progress over time can be preferred.

A number of options can be used to assess individual student growth for school accountability. For example, vertical scales, such as those based on item response theory (IRT) models (e.g., Roberts & Ma, 2006), are of interest. Vertical scales can be challenging to develop, their assumptions may not be met, and their promise may not be fulfilled (i.e., one scale score unit of growth may not have the same meaning at different parts of the scale and for different grades). Alternatives to vertical scales are regression-based approaches, such as empirical regression models (Yen, Lu, Smith, & Patz, 2006) and hierarchical regression models (e.g., Drury & Doran, 2003; Sanders & Horn, 1994).

This study explores the relationship between vertical scales and their alternatives in measuring academic progress for program evaluation using exemplar data from two content areas in Grades 4 to 10. The present study applies each of these models to the same data with the goal of arriving at a better understanding of the strengths and limitations of the models, and how they might differentially reflect growth.

Method

Data Source

To compare different methods of measuring academic growth in an empirical setting, data from the 2005 and 2006 administrations of the Qatar Comprehensive Educational Assessment (QCEA) were used. The QCEA is an annual series of standardized tests based on the Qatar Curriculum Standards for Arabic, English, mathematics, and science administered to students in Grades 4 to 11. These tests are developed under the auspices of the Student Assessment Office, which is part of the Evaluation Institute under Qatar's Supreme Education Council. The QCEA is an integral component of the educational reform efforts currently underway in Qatar. In addition to providing vital feedback to schools and teachers to improve learning, the results of the QCEA are an important component of the accountability framework for the Qatari Educational Reform. In this study, data from the Arabic and English QCEA assessments in grades 4 through 10 were used.

The Arabic and English assessments contain both multiple-choice and constructed-response items (47 to 61 items per test). These difficult tests reflect the challenging nature of the Qatar Curriculum Standards. In 2005 mean p-values ranged from 0.42 to 0.50 for Arabic and from 0.29 to 0.41 for English; in 2006 mean p-values ranged from 0.42 to 0.51 for Arabic and from 0.31 to 0.35 for English. In 2005 the reliabilities for the Arabic assessments ranged from 0.85 to 0.89 and from 0.76 to 0.90 for English. In 2006 the reliabilities were similar and ranged from 0.86 to 0.90 for Arabic and from 0.81 to 0.88 for the English assessments.

In 2005 IRT models were used to calibrate the test items at each grade level. The three-parameter logistic (3PL) model was used for multiple-choice items, and the generalized partial credit model (GPC; Muraki, 1992) was used for constructed-response items. In addition, cross-grade vertical scales were constructed separately for Arabic and English, applying the Stocking and Lord (1983) method to common items that were shared with the grade above (ETS, 2006).The validity of the vertical scale was supported by good model fit, the consistency of parameters across grades, and the progression of examinee scores across grade levels. That is, the mean scores improved with each increase in grade level, reflecting a steady increase in abilities as grade level increased. In addition, both assessments showed an overall ordinal progression in the test characteristic curves across grade levels, indicating that students with a given scale score had a lower expected proportion correct on a test at a higher grade level. The vertical scales for Arabic and English were maintained in 2006 by horizontal on-grade linking. In this study, scale scores for the Grades 4 to 10 tests ranged from 425 to 935 for Arabic and from 500 to 935 for English.[1]

Using unique student identification numbers, students' records were matched longitudinally from 2005 to 2006. For example, Grade 4 student records in 2005 were matched to their Grade 5 records in 2006. Tables 1 and 2 provide information on the matches for Arabic and English, respectively, across Grades 4 to 10. Matches were found for 62.0% to 80.3% of the valid cases. The number of matched cases ranged from 2940 (English Grades 9 to 10) to 5270 (Arabic Grades 4 to 5). Overall, the matched cases had mean scale scores within 1 to 3 scale score points of all valid cases, except for Arabic Grades 9 to 10 where the matched cases were 8 points higher than all valid cases.

Table 3 shows the number of schools at each grade transition as well as the 2006 summary sample size statistics. At these grades the 2006 sample size per school ranged from 1 to 349 students with mean sample size of 40 to 74 students.

Performance Levels

To assist with interpretation of test scores and provide a general idea of the progress the Qatari educational system is making in terms of improving student achievement in relation to the high goals set out by the Qatar Curriculum Standards, the Student Assessment Office adopted performance levels, consisting of descriptors and cut scores, for each subject area (ETS, 2006). Using a modified Angoff method (Hambleton & Plake, 1995), a panel of content experts set two cut scores distinguishing Below Standards from Approaches Standards and Approaches Standards from Meets Standards in Grades 5, 8, and 11. Using these recommended cut scores, cut scores for Grades 6, 7, 9, and 10 were interpolated, and cut scores for Grade 4 were extrapolated downward. A linear interpolation/extrapolation method in the metric of percentages of students scoring at or above the cut scores was used. The interpolation/ extrapolation procedures were designed to give full consideration to the judgments of the panelists; ensure that the results exhibited reasonable, consistent, and explainable cross-grade patterns in the percentage of students in each performance level, and remain consistent with the expectation that students in higher grades exhibit a greater degree of mastery than students in lower grades for any given performance level. This procedure was used on both the Arabic and English assessments.

Table 1

Scale Score Summary Statistics for Merged Records—Arabic

Grade match	2005 all valid cases			2005 matched records			2006 matched records			2006 all valid cases			Matched records	
	N	Mean SS	SD SS	*N*	Mean SS	SD SS	*N*	Mean SS	SD SS	*N*	Mean SS	SD SS	As % of all 2005	As % of all 2006
Grade 4 to Grade 5	6,559	466	35	5,270	467	35	5,270	478	39	6,575	476	40	80.3%	80.2%
Grade 5 to Grade 6	6,644	476	40	5,190	478	39	5,190	490	43	6,474	489	43	78.1%	80.2%
Grade 6 to Grade 7	6,206	489	42	4,821	490	41	4,821	499	43	6,553	498	45	77.7%	73.6%
Grade 7 to Grade 8	6,164	501	44	4,724	504	43	4,724	507	44	6,242	507	45	76.6%	75.7%
Grade 8 to Grade 9	5,557	508	45	4,044	511	44	4,044	520	42	5,450	520	43	72.8%	74.2%
Grade 9 to Grade 10	5,382	517	49	3,632	525	47	3,632	521	48	5,206	521	48	67.5%	69.8%

Note. The 2005 results are for the lower of the two grades being matched, and the 2006 results are for the higher grade. SS = scale score.

Table 2

Scale Score Summary Statistics for Merged Records—English

Grade match	2005 all valid cases			2005 matched records			2006 matched records			2006 all valid cases			Matched records	
	N	Mean SS	SD SS	*N*	Mean SS	SD SS	*N*	Mean SS	SD SS	*N*	Mean SS	SD SS	As % of all 2005	As % of all 2006
Grade 4 to Grade 5	6,072	508	14	4,707	508	14	4,707	514	18	6,394	514	18	77.5%	73.6%
Grade 5 to Grade 6	6,238	512	17	4,952	512	16	4,952	518	19	6,535	518	19	79.4%	75.8%
Grade 6 to Grade 7	5,769	516	18	4,495	516	18	4,495	522	20	6,420	521	20	77.9%	70.0%
Grade 7 to Grade 8	5,373	521	20	4,048	523	20	4,048	527	22	6,136	527	22	75.3%	66.0%
Grade 8 to Grade 9	4,927	526	23	3,632	527	23	3,632	530	24	5,336	529	24	73.7%	68.1%
Grade 9 to Grade 10	4,738	532	25	2,940	535	25	2,940	538	27	4,743	536	27	62.1%	62.0%

Note. The 2005 results are for the lower of the two grades being matched, and the 2006 results are for the higher grade. SS = scale score.

Table 3

Sample Size Summary for 2006 Schools by Grade

2006 grade	No. of 2006 schools	Mean sample size	SD of sample size	Minimum sample size	Maximum sample size
		Arabic			
Grade 5	123	43	31	1	203
Grade 6	95	55	55	1	349
Grade 7	66	73	53	1	208
Grade 8	64	74	53	1	178
Grade 9	64	63	49	2	209
Grade 10	52	70	49	4	195
		English			
Grade 5	115	40	29	3	167
Grade 6	95	52	52	1	346
Grade 7	63	71	51	1	166
Grade 8	63	64	50	1	198
Grade 9	67	54	44	1	184
Grade 10	53	55	42	1	168

In this study, two performance categories were used: Below Standards and at or above Approaches Standards. For the sake of simplified terminology, in this study the term Approaches Standards is taken to mean at or above Approaches Standards (i.e., Approaches Standards or Meets Standards).

The Approaches Standards cut scores for Grades 4 to 10 are presented in Table 4. Tables 5 and 6 provide performance level information for the matched sample for Arabic and English, respectively, across Grades 4 to 10. In most cases, the percentage of students in the Approaches Standards performance level is very similar from 2005 to 2006. In addition, the percentages are low, particularly in English, reflecting the challenging nature of the Qatar Curriculum Standards and performance levels.

The correlations between scale scores and the dichotomous proficiency level categories are presented in Table 7. The correlations between scale scores in 2005 and 2006 ranged from 0.68 to 0.74 in Arabic and from 0.59 to 0.73 in English. The correlations between students' scale scores in 2005 and their proficiency levels in 2006 ranged from 0.54 to 0.62 in Arabic and from 0.51 to 0.67 in English. The cross-grade correlations of the proficiency levels ranged from 0.56 to 0.60 in Arabic and from 0.58 to 0.64 in English. The fact that these assessments were difficult for the Qatari students is likely to have attenuated these correlations to some extent.

Table 4

Approaches Standards Cut Scores for Arabic and English

Grade	Approaches Standards cut score	
	Arabic	English
Grade 4	493	530
Grade 5	505	537
Grade 6	521	541
Grade 7	533	549
Grade 8	541	557
Grade 9	553	564
Grade 10	555	570

Academic Growth Measures

In addition to the vertical scale, several alternatives for evaluating growth that do not require vertical scales were examined.[2] Growth was measured in scale score units, as well as by changes in performance levels. The methods used to measure growth, which we have chosen to generically describe as projections, are listed below, and they are described in detail in the following section.

- Vertical scale
  - Students' scale score differences across grade levels
  - Students' performance level differences across grade levels
- Empirical cross-grade regressions
  - Regression of scale score on the students' scale scores from the previous grade
  - Regression of the dichotomous variable Below Standards/ Approaches Standards on the students' scale scores from the previous grade
- Ordinary least squares (OLS) and logistic regressions
  - Linear regression of scale score on the students' scale scores from the previous grade
  - Logistic regression of the dichotomous variable Below Standards/Approaches Standards on the students' scale scores from the previous grade

Table 5

Performance Level Summary Statistics for Merged Records—Arabic

	2005 all valid cases		2005 matched records		2006 matched records		2006 all valid cases	
Grade match	*N*	Approaches Standards	*N*	Approaches Standards	*N*	Approaches Standards	*N*	Approaches Standards
Grade 4 to Grade 5	6,559	23%	5,270	25%	5,270	29%	6,575	27%
Grade 5 to Grade 6	6,644	26%	5,190	27%	5,190	26%	6,474	26%
Grade 6 to Grade 7	6,206	25%	4,821	25%	4,821	25%	6,553	24%
Grade 7 to Grade 8	6,164	27%	4,724	28%	4,724	26%	6,242	27%
Grade 8 to Grade 9	5,557	27%	4,044	29%	4,044	22%	5,450	23%
Grade 9 to Grade 10	5,382	27%	3,632	31%	3,632	27%	5,206	26%

Note. The 2005 results are for the lower of the two grades being matched, and the 2006 results are for the higher grade.

Table 6

Performance Level Summary Statistics for Merged Records—English

	2005 all valid cases		2005 matched records		2006 matched records		2006 all valid cases	
Grade match	*N*	Approaches Standards	*N*	Approaches Standards	*N*	Approaches Standards	*N*	Approaches Standards
Grade 4 to Grade 5	6,072	9%	4,707	9%	4,707	13%	6,394	13%
Grade 5 to Grade 6	6,238	10%	4,952	11%	4,952	14%	6,535	14%
Grade 6 to Grade 7	5,769	10%	4,495	10%	4,495	11%	6,420	11%
Grade 7 to Grade 8	5,373	10%	4,048	11%	4,048	10%	6,136	10%
Grade 8 to Grade 9	4,927	11%	3,632	11%	3,632	9%	5,336	9%
Grade 9 to Grade 10	4,738	11%	2,940	13%	2,940	13%	4,743	12%

Note. The 2005 results are for the lower of the two grades being matched, and the 2006 results are for the higher grade.

Table 7

Cross-Grade Correlations of Student-Level Scores for Arabic and English

2005–2006 grades	*N*	Correlation Between 2005 SS and 2006 SS	 Between 2005 SS and 2006 PL	 Between 2005 PL and 2006 PL
		Arabic		
Grade 4 to Grade 5	5,270	0.73	0.62	0.58
Grade 5 to Grade 6	5,190	0.74	0.61	0.60
Grade 6 to Grade 7	4,821	0.72	0.58	0.57
Grade 7 to Grade 8	4,724	0.70	0.56	0.56
Grade 8 to Grade 9	4,044	0.70	0.54	0.57
Grade 9 to Grade 10	3,632	0.68	0.54	0.58
		English		
Grade 4 to Grade 5	4,707	0.73	0.67	0.60
Grade 5 to Grade 6	4,952	0.70	0.62	0.58
Grade 6 to Grade 7	4,495	0.59	0.53	0.60
Grade 7 to Grade 8	4,048	0.59	0.53	0.61
Grade 8 to Grade 9	3,632	0.61	0.51	0.64
Grade 9 to Grade 10	2,940	0.59	0.52	0.62

Note. PL = proficiency level, SS = score scale.

- Multilevel analyses
  - Multilevel linear regression of scale score on the students’ scale scores from the previous grade, which takes school context into account by modeling the clustering of students nested within schools
  - Multilevel logistic regression of the dichotomous variable Below Standards/Approaches Standards on the students’ scale scores from the previous grade, which takes school context into account by modeling the clustering of students nested within schools

Vertical scale. As described in the Data Source section, vertical scales across Grades 4 to 10 were developed in 2005 for the Arabic and English tests, respectively, and maintained by horizontal on-grade linking in 2006. Using the vertical scales across Grades 4 to 10, scale scores can be compared across grades. Thus, growth is assessed by looking at the change in students' scale scores and performance levels in Arabic and English from one grade to the next.

Empirical regression. The empirical regression involves calculating the conditional means of the 2006 scale scores at each of the 2005 scale scores. In the empirical regression, results were restricted to be monotonic nondecreasing. That is, the empirical 2006 conditional means were smoothed to be strictly nondecreasing as the 2005 score increased.[3] In order to preserve the empirical results as much as possible, while assuring monotonicity, the pooled adjacent violators algorithm (PAVA; Barlow, Bartholomew, Bremner, & Brunk, 1972) adapted from the implementation of Raubertas (1994) was applied. The algorithm iteratively identifies and replaces nonmonotone 2006 conditional means by their weighted (by the number of students) averages of adjacent values until the 2006 conditional means were strictly nondecreasing as the 2005 score increased. The 2006 conditional means that were monotone in the unsmoothed data remained unchanged by the application of the PAVA.

For the proficiency levels, the conditional percentages of Approaches Standards students in 2006 were obtained for each of the 2005 scale scores. The PAVA was also used to produce monotonicity for the percentage of students in the Approaches Standards category.

Ordinary least squares and logistic regressions. In addition, an OLS regression model was used to obtain the projections for the 2006 scale scores:

$$Y_i = \beta_0 + \beta_1 (X_i - \overline{X}) + \varepsilon_i \, , \tag{1}$$

where Y_i is the 2006 scale score for student i , X_i is the 2005 scale score for student *i*, $\overline{X}$ represents the mean 2005 scale score taken over students (i.e., the grand mean), β_0 and β_1 are the model parameters, and ε_i is the random error term.

Because the proficiency level categorization Below Standards versus Approaches Standards is a dichotomous variable, a logistic regression was used to obtain the 2006 proficiency level projections. The specific form of the logistic model used is as follows:

$$P_i = \frac{\exp(\beta_0 + \beta_1 (X_i - \overline{X}))}{1 + \exp(\beta_0 + \beta_1 (X_i - \overline{X}))} \, , \tag{2}$$

where P_i is the probability of a 1 (the probability of Approaches Standards) for student *i*, β_0 and β_1 are the model parameters, X_i is the 2005 scale score for student *i*, and $\bar{X}$ represents the mean 2005 scale score over students (i.e., the grand mean).

Multilevel models. The multilevel models used in this study extend the OLS and logistic regression approaches described above by taking into account the nesting of students within schools. Multilevel models are useful when the data being analyzed provide sufficient variation within and between levels (Heck & Thomas, 2000). An unconditional random intercepts model was used as a baseline to assess how variation in the 2006 outcome is allocated across the levels of analysis (Bryk & Raudenbush, 1992). The 2005 scale scores, centered at the grand (country) mean, were then entered as predictors at the student level. Exploratory analysis indicated that the slopes did not vary significantly over schools, therefore the slope was held fixed across schools.[4]

In the following sections, we describe in more detail the models used in this study.

Multilevel linear models for 2006 scale score. Because the model reflects the nesting of students within schools, we begin by describing the student model and then progress to the school model.

Level I, or student model. Given student *i* attending school *j*, the student level regression model is given by:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}, \tag{3}$$

where Y_{ij} is the 2006 scale score for student *i* in school *j*, X_{ij} is the 2005 scale score for student *i* in school *j*, and $\bar{X}_{..}$ represents the mean 2005 scale score over students or grand mean. The intercept, β_{0j}, is the expected 2006 score for a student in school *j* whose 2005 score is equal to the grand mean; β_{1j} is the regression slope for 2006 scores on 2005 scores; and ε_{ij} is the random error associated with the *i*th student in the *j*th school. The student-level random errors are assumed to be normally distributed with a mean of 0 and constant variance of σ^2.

Level II, or school model. At the school level, the Level I regression coefficients are modeled as:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{4}$$

$$\beta_{1j} = \gamma_{01}, \tag{5}$$

where γ_{00} is the average of the school means on 2006 scale score across the population of schools, u_{0j} is increment in the intercept attributable to school *j* or the school random intercept effect, and γ_{01} represents the pooled within-school regression slope for 2006 scores on 2005 scores. The school random intercept effects, u_{0j}, are assumed to be independently distributed $\sim N(0, \tau_{00})$, where τ_{00} is the variance of school intercepts. After substitution the combined model is given by

$$Y_{ij} = \gamma_{00} + \gamma_{01}(X_{ij} - \bar{X}_{..}) + u_{0j} + \varepsilon_{ij}, \tag{6}$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$.

In multilevel analysis for school accountability, we are particularly interested in the school intercepts, u_{0j}—which are the variables typically used in analyzing school effects in value-added models—and size of their variance, τ_{00}. If the variance of the school intercepts is large relative to the total variation, there is substantial evidence of between-school differences in the outcome of interest, suggesting measurable school effects on students' test scores. A relatively large school intercept variation would also suggest that an OLS regression analysis is inappropriate for the data and would likely yield results that are misleading at the school level. Conversely if the variation in school intercepts is small, the results for OLS and multilevel regressions would match. The measure used in this study of the relative size of the between-school variation is the intraclass correlation coefficient:

$$\hat{\rho} = \hat{\tau}_{00} \big/ (\hat{\tau}_{00} + \hat{\sigma}^2). \tag{7}$$

If calculated after entering explanatory or predictor variables, it is referred to as an adjusted intraclass correlation coefficient. This coefficient can also be thought of as the correlation between pairs of 2006 scores in the same school after adjusting for the 2005 score.

It is important to note that in multilevel modeling, the unknown true mean of each school *j*, β_{0j}, is calculated as an empirical Bayes (EB) estimator (Lindley & Smith, 1972) or weighted combination of the observed school mean, $\bar{Y}_{.j}$, and the estimated average of the school means, $\hat{\gamma}_{00}$. The EB estimate β^{*}_{0j} is given by:

$$\beta^{*}_{0j} = \lambda_j \bar{Y}_{.j} + (1 - \lambda_j)\hat{\gamma}_{00} \tag{8}$$

where λ_j is the reliability of $\bar{Y}_{.j}$ for the parameter β_{0j} (Kelley, 1927). The EB estimate lies between the school mean and the grand mean. If the reliability is low, Equation 8 indicates that the EB estimate for the school mean will shrink toward the estimated grand mean. The reliability of the individual school intercept estimate is the ratio of the parameter variance relative to the total variance of the sample mean given by:

$$\lambda_j = \frac{Var(\beta_{0j})}{Var(\bar{Y}_{.j})} = \frac{\tau_{00}}{(\tau_{00} + \sigma^2 / n_j)}\,. \tag{9}$$

This reliability will be high when there is substantial variation among schools, holding sample size constant, or when the number of students per school is large. In the data set used for this study, there are a number of schools with small sample size, and the reliability estimate will be used to assess the shrinkage of the EB estimates for the school means.

Multilevel nonlinear models for 2006 proficiency level. The proficiency level of student *i* in school *j*, Y_{ij}, is distributed as a Bernoulli random variable with expected value $E(y_{ij} \mid p_{ij}) = p_{ij}$ and variance $Var(y_{ij} \mid p_{ij}) = p_{ij}(1 - p_{ij})$, where p_{ij} is the probability that student *i* in school *j* is in the Approaches Standards performance category. For the student-level model, the logit link function is used to linearize the relationship between the log odds of Approaches Standards and the structural or predictor variables. The school-level model takes the same form as in the linear model above. The combined model is given by:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \gamma_{00} + \gamma_{01}(X_{ij} - \bar{X}_{..}) + u_{0j}\,, \tag{10}$$

where γ_{00} represents the adjusted mean log-odds of classification in the Approaches Standards performance level over schools, u_{0j} is the school random intercept effect, and γ_{01} represents the pooled within-school slope. As with the linear model, the school random intercept effects are assumed independently distributed.

Unlike the linear model applied to scale scores, the intraclass correlation coefficient is not particularly informative for the nonlinear model applied to performance levels, because the Level I variance is heteroscedastic (Raudenbush & Bryk, 2002). Therefore, the intraclass correlation was not calculated for the nonlinear model.

As a measure of fit, an extradispersion parameter was obtained. This parameter measures the extent to which the observed errors follow the theoretical binomial error distribution. The extradispersion parameter should be close to 1.[5] Values greater than 1 indicate greater variability in the Level I outcome than expected under the Level I sampling model, while values substantially less than 1 indicate less variability than expected.

Results

Empirical Regression

The conditional means (empirical regressions) of the 2006 scale scores at each 2005 scale score for Grades 4 to 10 are shown graphically in Figures 1 and 2 for Arabic and English, respectively. Viewed globally, the empirical regressions are fairly linear through much of the distribution but less linear in the lower tails, where the regressions tend to flatten out. The empirical regressions are fairly similar across grades, but not identical. Figures 3 and 4 present the conditional percentages of Approaches Standards students in 2006 at each 2005 scale score point for Arabic and English. The progression of these graphs over grades demonstrates the vertical moderation of the performance standards.

Ordinary Least Squares and Logistic Regression

Table 8 presents the results of fitting OLS regression models to the scale score data from Grades 4 to 10 for Arabic and English. The table displays OLS estimated intercepts and slopes along with the associated standard errors, root mean square error (RMSE), and R^2 statistics. The estimated model parameters and their associated standard errors show that the 2005 scale score is a statistically significant predictor of the 2006 scale score at all grade levels for both content areas. For Arabic, the RMSE ranged from 26.7 to 35.2, and for English, the values ranged from 12.2 to 21.8. The R^2 are moderate with values from 46% to 54% for Arabic and from 35% to

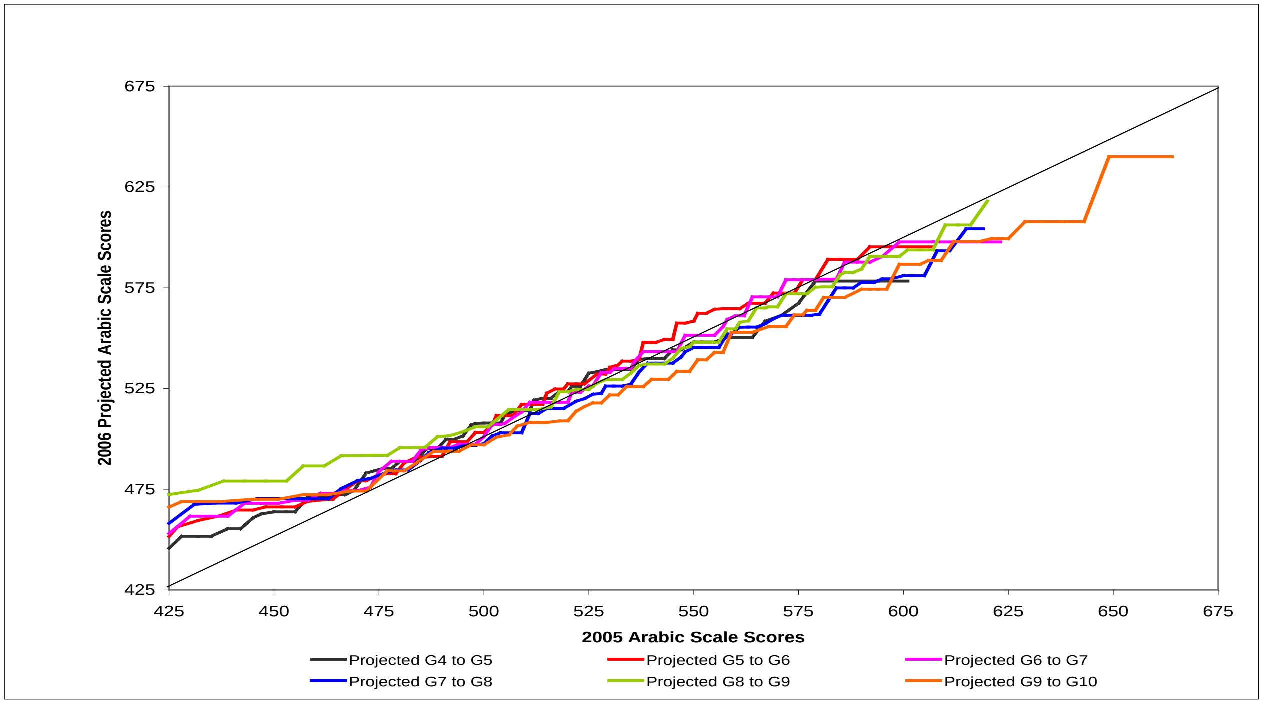

Figure 1. **Empirical regression: 2006 projected Arabic scale scores versus 2005 Arabic scale scores.**

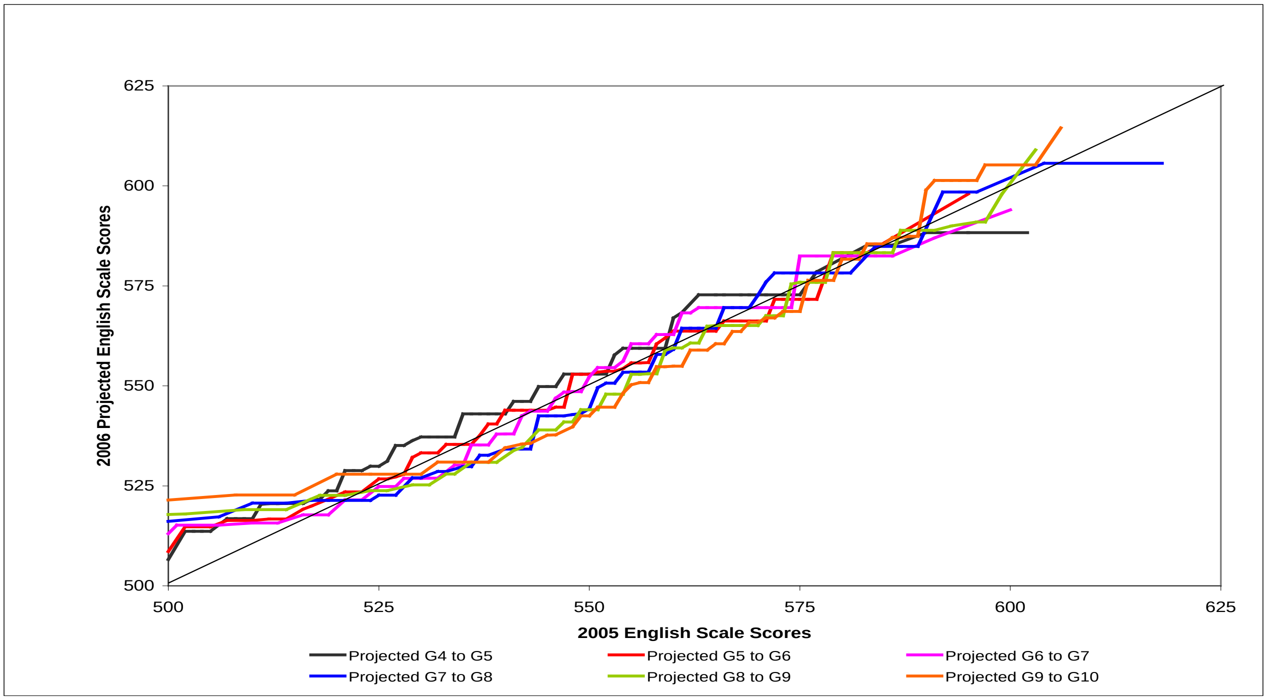

Figure 2. **Empirical regression: 2006 projected English scale scores versus 2005 English scale scores.**

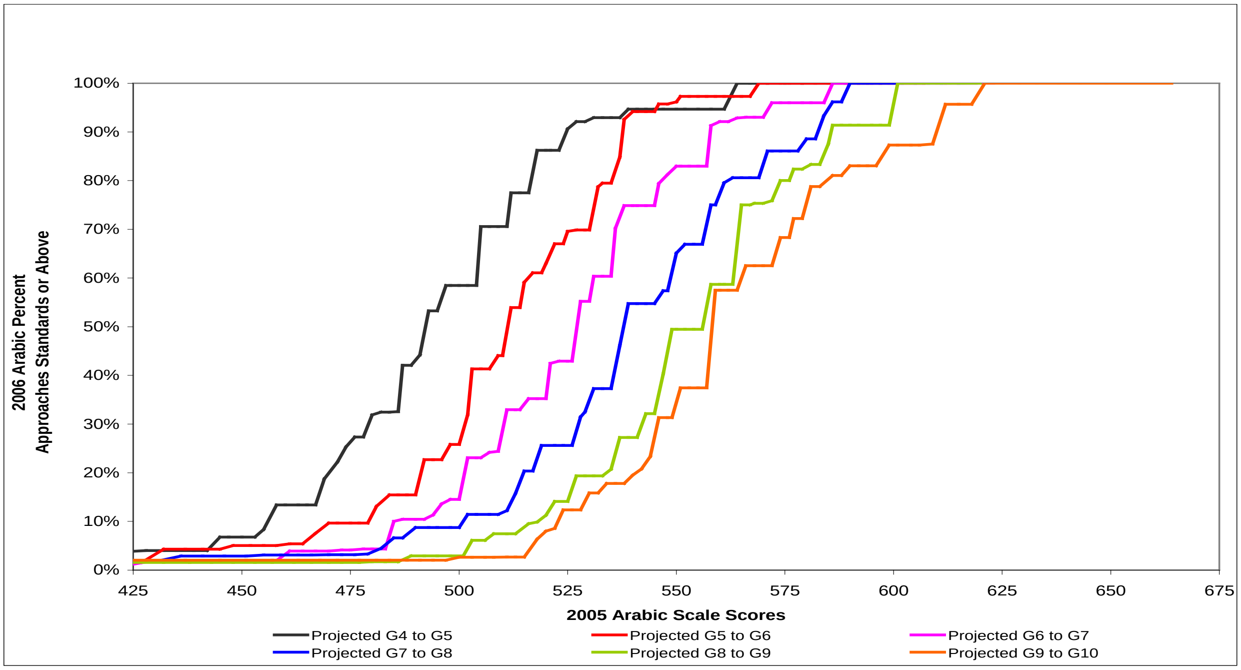

Figure 3. **Empirical regression: 2006 projected Arabic percentage Approaches Standards versus 2005 Arabic scale scores.**

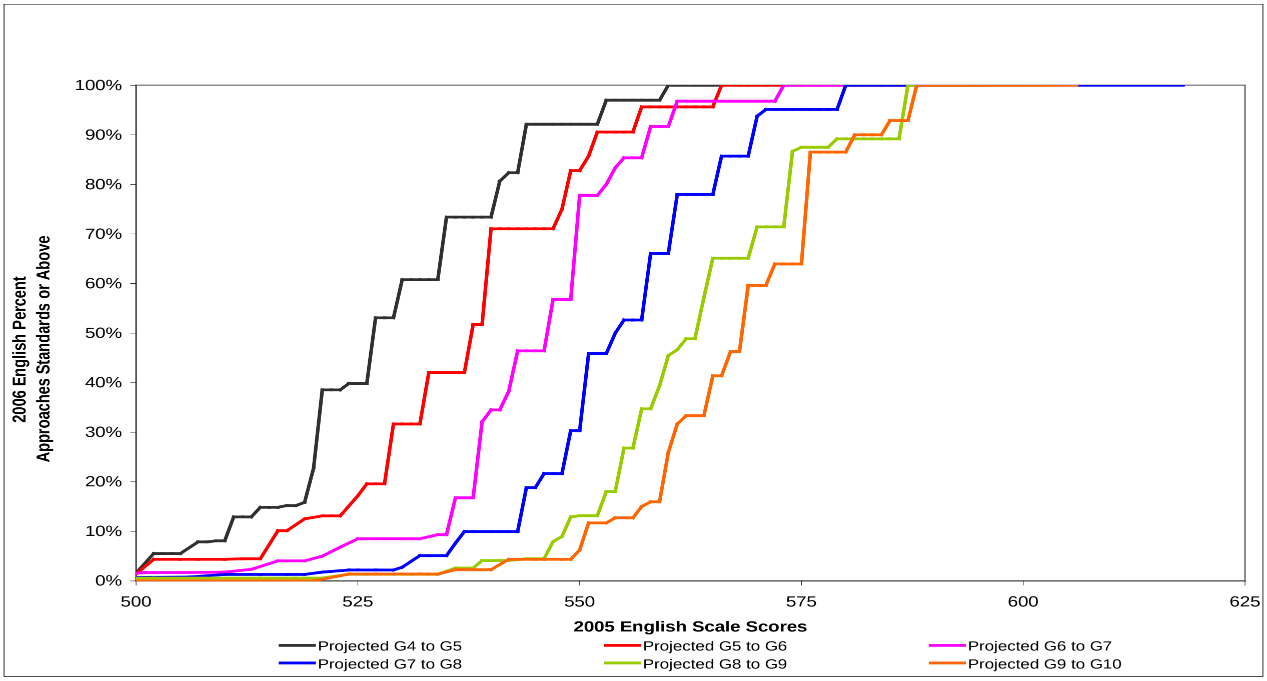

Figure 4. **Empirical regression: 2006 projected English percentage Approaches Standards versus 2005 English scale scores.**

Table 8

Results of Fitting Ordinary Least Squares Regression Models to Project 2006 Scale Score From 2005 Scale Scores

Projection	N	Intercept		Slope		RMSE	R^2
		Estimate	SE	Estimate	SE		
				Arabic			
Grade 4 to Grade 5	5,270	478.3	0.4	0.82	0.01	26.7	0.54
Grade 5 to Grade 6	5,190	489.8	0.4	0.81	0.01	28.8	0.54
Grade 6 to Grade 7	4,821	499.1	0.4	0.79	0.01	30.6	0.52
Grade 7 to Grade 8	4,724	507.3	0.5	0.72	0.01	31.8	0.49
Grade 8 to Grade 9	4,044	519.6	0.5	0.68	0.01	30.1	0.49
Grade 9 to Grade 10	3,632	521.2	0.6	0.70	0.01	35.2	0.46
				English			
Grade 4 to Grade 5	4,707	514.5	0.2	0.95	0.01	12.2	0.53
Grade 5 to Grade 6	4,952	518.4	0.2	0.81	0.01	13.7	0.49
Grade 6 to Grade 7	4,495	521.7	0.2	0.68	0.01	16.2	0.35
Grade 7 to Grade 8	4,048	526.8	0.3	0.64	0.01	17.6	0.35
Grade 8 to Grade 9	3,632	530.5	0.3	0.64	0.01	19.1	0.37
Grade 9 to Grade 10	2,940	537.9	0.4	0.64	0.02	21.8	0.35

53% for English. The Arabic cross-grade scale score correlations are very consistent from Grades 4 to 7, are a bit lower for Grades 8 and 9, and are a bit lower still for Grade 10. The English cross-grade scale score correlations decrease markedly from Grades 5 and 6 to the higher grades.

As an example, the 2006 Arabic Grade 5 projected scale score for a student with an Arabic Grade 4 scale score of 453 is given by

$$\hat{Y} = 478.3 + 0.82(453 - 467) = 467\,, \tag{11}$$

where 478.3 is the intercept parameter from Table 8, and 0.82 is the slope parameter as show in Table 8 and 467 is the grand mean for the 2005 Grade 4 scale scores. Figures 5 and 6 show for Arabic and English the projected 2006 scale score at each of the 2005 scale scores based on the linear regressions. These plots show that both the fitted intercepts and slopes are fairly consistent across the grades for Arabic, reflecting the homogeneity observed in the empirical regression projections. The slopes for the English assessments were more heterogeneous than the slopes for the Arabic assessments.

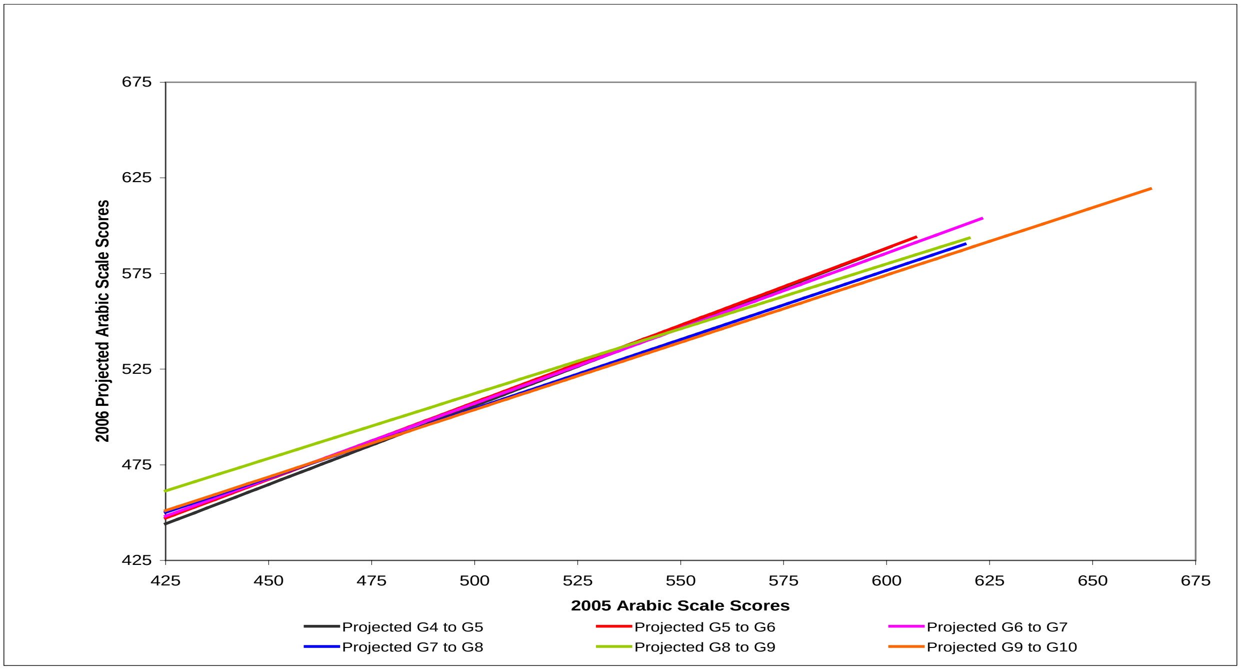

Figure 5. **Ordinary least square regression: 2006 projected Arabic scale scores versus 2005 Arabic scale scores.**

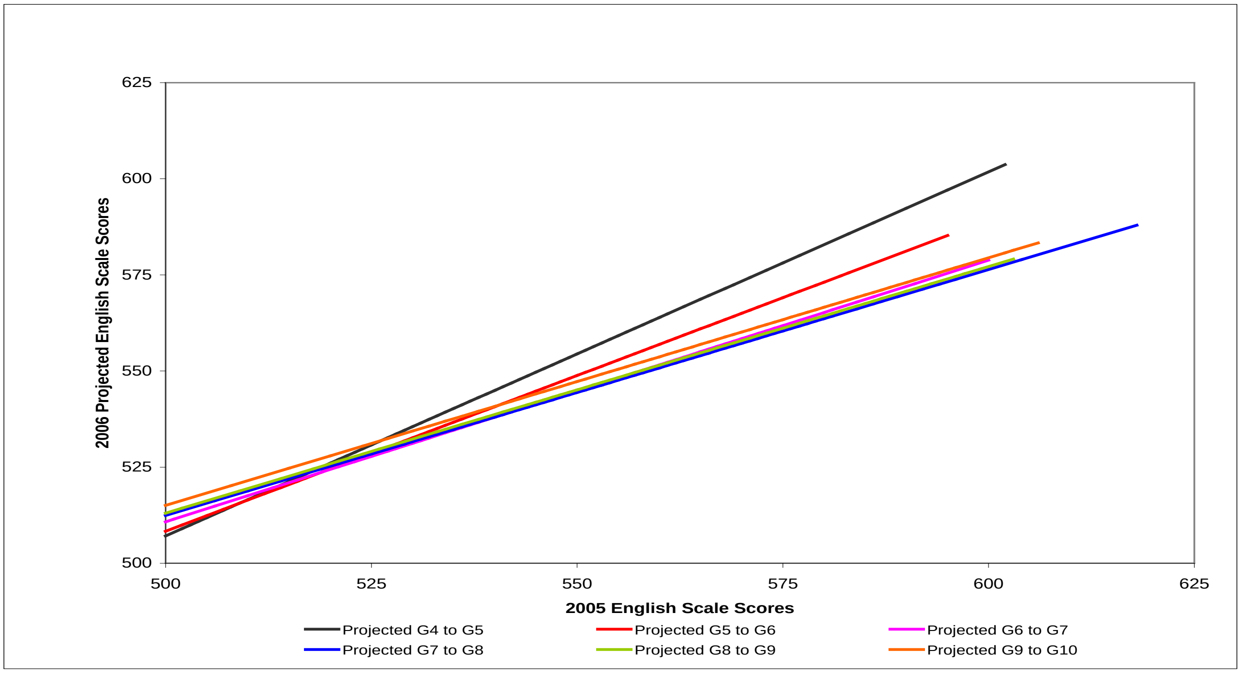

Figure 6. **Ordinary least square regression: 2006 projected English scale scores versus 2005 English scale scores.**

The OLS regressions in Figures 5 and 6 can be compared with the empirical regressions in Figures 1 and 2. (See also Figures 13 and 14.) The OLS regressions appear to do a good to fair job of representing all the Arabic empirical regressions and the English empirical regressions at Grades 5 and 6, but a poor job for English Grades 7 to 10. The linear OLS regressions fit the empirical results better at the lower end of the scale and do not wholly capture the curvilinearity of the empirical regressions. In evaluating these regressions, it is essential to keep in mind that most students are near the lower end of the scale and that there are few students at the high end of the scale where the largest differences between the methods occur.

The results for the logistic regression for the performance levels are presented in Table 9 for all grades and content areas. Table 9 contains the estimates of the intercept and slopes, as well as the standard errors of the estimated coefficients. Also presented in Table 9 is the odds ratio, which shows the odds of being in the Approaches Standards performance level per unit increase in 2005 scale score above the grand mean. For example, the value of 1.06 for Arabic Grade 4 indicates that the odds of being in the Approaches Standards performance level increases by 6% for every 1 point increase in scale score above the grand mean. The χ^2 statistic (1 d.f.) for the likelihood ratio test comparing the intercept-only model versus the logistic model with an intercept and the 2005 scale score as a covariate is highly statistically significant at all grade levels and content areas, indicating that the 2005 scale score should be used in projecting the 2006 performance level.

As an example of model-based projection for the student with the Arabic 2005 grade 4 scale score of 453, the probability of being in the Approaches Standards performance level is given by

$$\hat{P} = \frac{\exp(-1.5 + 0.05(453 - 467))}{1 + \exp(-1.5 + 0.05(453 - 467))} = 0.098, \qquad (12)$$

where -0.15 and 0.05 are the model parameter estimates from Table 9 and 467 is the grand mean for the 2005 Grade 4 scale scores.

The projected percentage of students in the Approaches Standards performance level based on the logistic regression parameter estimates are shown in Figures 7 and 8 for Arabic and English, respectively. As with the empirical regressions, the logistic projection curves show a progressive ordering across the grade levels for both content areas. Figures 7 and 8 can be

Table 9

Results of Fitting Logistic Regression Models to Project 2006 Performance Level From 2005 Scale Scores

Projection	Intercept		Slope		Odds ratio	χ^2
	Estimate	SE	Estimate	SE		
Arabic						
Grade 4 to Grade 5	-1.5	0.05	0.05	0.002	1.06	2267
Grade 5 to Grade 6	-1.9	0.06	0.06	0.002	1.06	2293
Grade 6 to Grade 7	-2.0	0.06	0.06	0.002	1.06	2029
Grade 7 to Grade 8	-1.8	0.06	0.05	0.002	1.05	1875
Grade 8 to Grade 9	-2.4	0.08	0.05	0.002	1.06	1585
Grade 9 to Grade 10	-2.0	0.07	0.05	0.002	1.06	1504
English						
Grade 4 to Grade 5	-2.8	0.08	0.13	0.005	1.14	1672
Grade 5 to Grade 6	-2.8	0.08	0.11	0.004	1.12	1708
Grade 6 to Grade 7	-3.4	0.10	0.11	0.004	1.12	1212
Grade 7 to Grade 8	-4.0	0.14	0.12	0.005	1.13	1176
Grade 8 to Grade 9	-4.5	0.18	0.12	0.006	1.13	1080
Grade 9 to Grade 10	-4.3	0.18	0.13	0.007	1.14	1087

compared with the empirical regressions in Figures 3 and 4. (See also Figures 15 and 16.) The logistic regressions appear to do a good to fair job of representing the empirical regressions at the lower scale scores where most of the students are scoring.

Multilevel Models

Table 10 presents the results of fitting multilevel linear models to the data from Grades 4 to 10 for Arabic and English. Presented are the fixed effects for the intercept (γ_{00}) and slope (γ_{01}), the school- and student-level variance components (τ_{00} and σ^2, respectively), the adjusted intraclass correlations (ρ^*), and the mean λ_j over schools (reliability of $\overline{Y}_{.j}$ in estimating the parameter β_{0j}). A review of Table 10 shows that in both content areas, the intercepts approximate the observed grand means. The fixed effects for the slopes and their associated standard errors indicate that the 2005 scale score is a statistically significant predictor of the 2006 scale score at all grade levels for both content areas. The size of the school-intercept

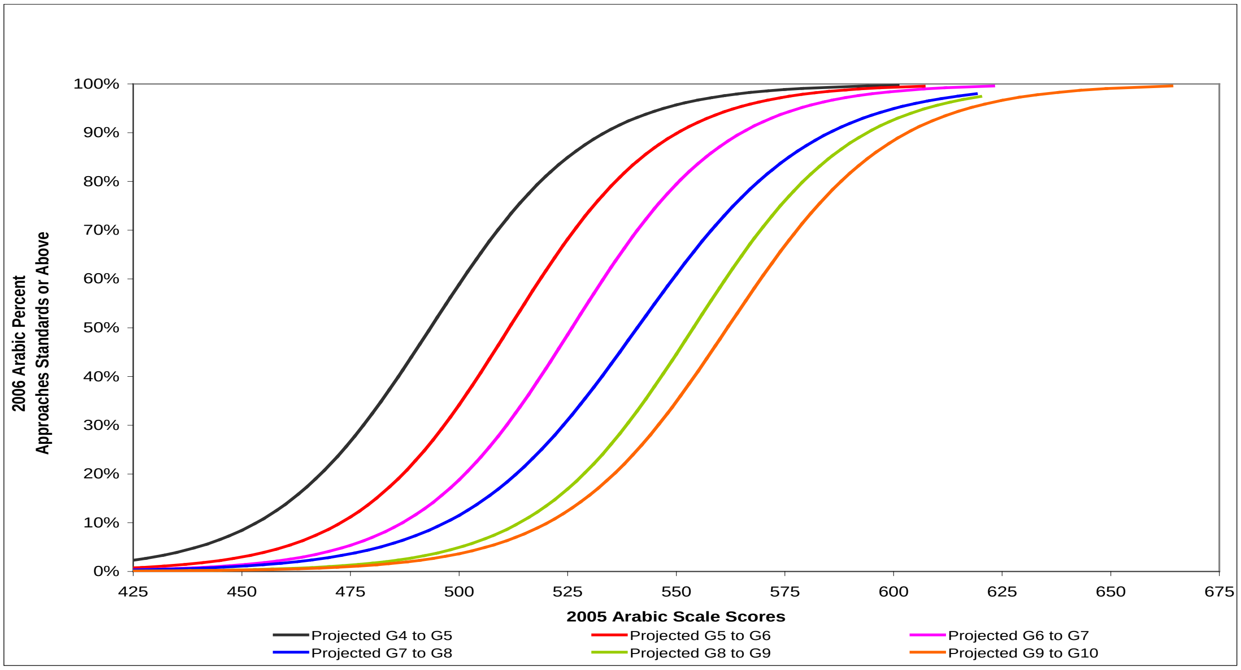

***Figure 7*. Logistic regression: 2006 projected Arabic percentage proficient versus 2005 Arabic scale scores.**

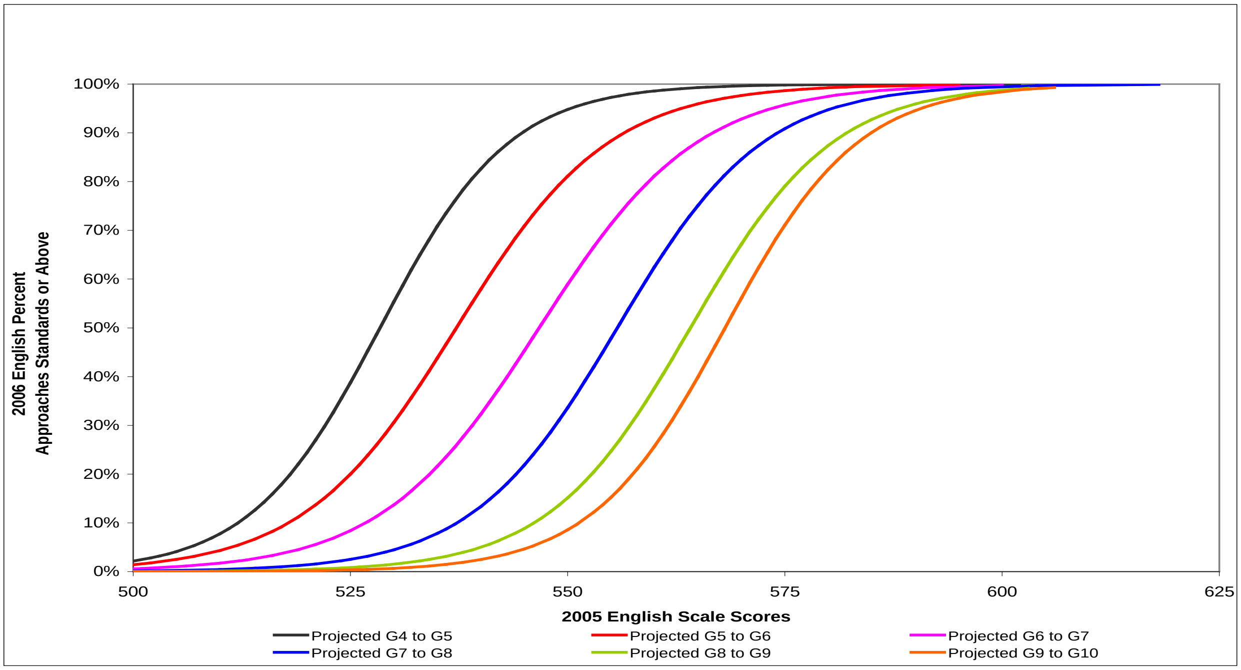

***Figure 8*. Logistic regression: 2006 projected English percentage proficient versus 2005 English scale scores.**

Table 10

Multilevel Linear Model Results for Scale Scores

Fixed effect	Grade 4 to Grade 5	Grade 5 to Grade 6	Grade 6 to Grade 7	Grade 7 to Grade 8	Grade 8 to Grade 9	Grade 9 to Grade 10
			Arabic			
Intercept, γ_{00}	478.0 (0.9)	490.6 (1.1)	498.2 (1.4)	507.4 (1.6)	518.1 (1.7)	519.0 (2.5)
Slope, γ_{01}	0.78 (0.01)	0.74 (0.01)	0.72 (0.01)	0.66 (0.01)	0.62 (0.01)	0.59 (0.01)
Variance component						
School intercept, τ_{00}	74.8 (12.6)	84.5 (17.9)	103.5 (21.7)	137.0 (29.4)	150.6 (33.3)	281.2 (62.2)
Level I error, σ^2	646.6	765.4	834.5	887.6	804.9	1,006.0
Intraclass correlation[a], ρ	0.10	0.10	0.11	0.13	0.16	0.22
Mean λ_j	0.75	0.74	0.81	0.84	0.86	0.90
			English			
Intercept, γ_{00}	513.7 (0.5)	518.3 (0.6)	521.2 (1.0)	526.0 (0.9)	530.0 (0.8)	536.1 (1.4)
Slope, γ_{01}	0.79 (0.02)	0.71 (0.01)	0.51 (0.01)	0.53 (0.01)	0.56 (0.02)	0.47 (0.02)
Variance component						
School intercept, τ_{00}	19.1 (3.1)	25.6(4.7)	55.8 (11.2)	37.6 (8.4)	27.3 (6.5)	93.9 (21.0)
Level I error, σ^2	131.1	165.3	214.3	277.8	339.7	389.2
Intraclass correlation [a], ρ	0.13	0.13	0.21	0.12	0.07	0.19
Mean λ_j	0.79	0.78	0.88	0.80	0.70	0.86

Note. Standard errors of the intercepts and slopes appear in parentheses under those values.

[a] The adjusted intraclass correlation is the proportion of variation that remains between schools after accounting for 2005 scale scores.

variance components, τ_{00}, relative to their standard errors indicate there is substantial variation in mean school performance. The mean λ_j is provided as an indicator of the shrinkage to be expected in schools with small sample sizes. Mean reliabilities range from 0.90 (Arabic Grades 9 to 10) to 0.70 (English Grades 8 to 9), indicating varying degrees of shrinkage toward the grand mean for some school estimates.[6]

For Arabic, the gradual decrease of the slopes indicates that the association between 2005 and 2006 scale scores weakens towards the upper grades. The adjusted intraclass correlations increase with grade ranging from approximately 0.10 at the lower grades to 0.22 for Grade 10, indicating that after controlling for school differences in 2005 scores, as much as 22% of the variation in 2006 scores is attributable to clustering of students within schools. The mean reliabilities were satisfactory, and ranged from 0.74 (Grade 5) to 0.90 (Grade 10), increasing with grades. Thus, shrinkage toward the mean in estimates of school effects was greater in the lower grade transitions than the upper grade transitions.

The results for English are somewhat different from Arabic. In general, the variance component for the school intercept, τ_{00} and the Level I error variance, σ^2, are relatively small compared to the Arabic tests, and the size of the slope and the adjusted intra-class correlation coefficients vary somewhat across grades. For example, the association (slope) between 2005 and 2006 scale scores decreases after Grade 5. The adjusted intraclass correlation coefficient is smallest in grades 8 to 9 (0.07) and largest in Grades 6 to 7 (0.21). The mean reliabilities were satisfactory and ranged from 0.70 (Grade 8 to 9) to 0.88 (Grade 6 to 7). As shown previously in Equation 6, the projected 2006 scale score for student *i* in school *j* is determined by the fixed regression coefficients as well as the school *j* random intercept effect. For example, using the parameter estimates in Table 10, the projected 2006 Arabic Grade 5 scale score for a student with an Arabic Grade 4 scale score of 453 would differ by school, according to the school's random intercept effect. If attending a school with a random intercept effect of 12.4, the projected 2006 Arabic scale score is calculated as

$$\hat{Y} = 478 + 0.78(453 - 467) + 12.4 = 479, \qquad (13)$$

where 478 is the fixed component of the intercept parameter estimate, γ_{00}, that represents the estimated average of the school means on the 2006 scale score across all schools; 0.78 is the fixed slope parameter, γ_{01}, that represents the expected increase in 2005 scale score for every point above the grand mean; 467 is the 2005 Grade 4 grand mean; and 12.4 is the incremental increase in the intercept attributable to the school attended by the student.

Figures 9 and 10 show the Arabic and English multilevel model projected 2006 scale score at each of the 2005 scale score points. These projections are based only on the fixed part of the model and assume that a student will achieve the average schooling experience in country—that is, will attend a school with effect 0. Individual schools will have projection lines that deviate from the overall projection for the country by the school-specific increment in the intercept. The plots show that both the fitted intercepts and slopes are fairly consistent with the results seen in the OLS regression, although for each grade, the fitted lines for the multilevel model are flatter due to shrinkage of fitted values towards the grand mean. The slopes for the English assessments differ more across grades than do those for the Arabic assessments, and the Grade 4 to 5 prediction line is noticeable for the steepness of its slope.

Table 11 presents the results of fitting multilevel logistic models to the performance classification data from Grades 4 to 10 for Arabic and English, respectively. All model parameters are given in the logit metric. The fixed effects for the slopes and their associated standard errors indicate that the 2005 scale score is a statistically significant predictor of the 2006 performance level in both content areas at all grade levels. However, inspection of the variance components and fit statistics indicate poor model-data fit for some of the Arabic and English grade transitions. As a measure of fit, the extradispersion parameter should be close to 1. Values greater than 1 indicate greater variability in the Level I outcome than expected under the level I sampling model, while values substantially less than 1 indicate less variability than expected. For example, the extradispersion parameter for Arabic Grades 8 to 9 is 3.24 and English Grades 8 to 9 is 5.49, indicating there may be other explanatory variables beyond previous year scale score that predict students' performance level within schools. Mild underdispersion appears in many of the grades indicating less variability (restriction of range) among student outcomes than expected. For both Arabic and English, the mean reliabilities (λ_i) were low, and ranged from 0.00 (English Grades 8 to 9) to 0.62 (English Grades 6 to 7), reflecting in part a restriction of range in the percentages Approaches Standards, especially in the Grade 8 to 9 transition.

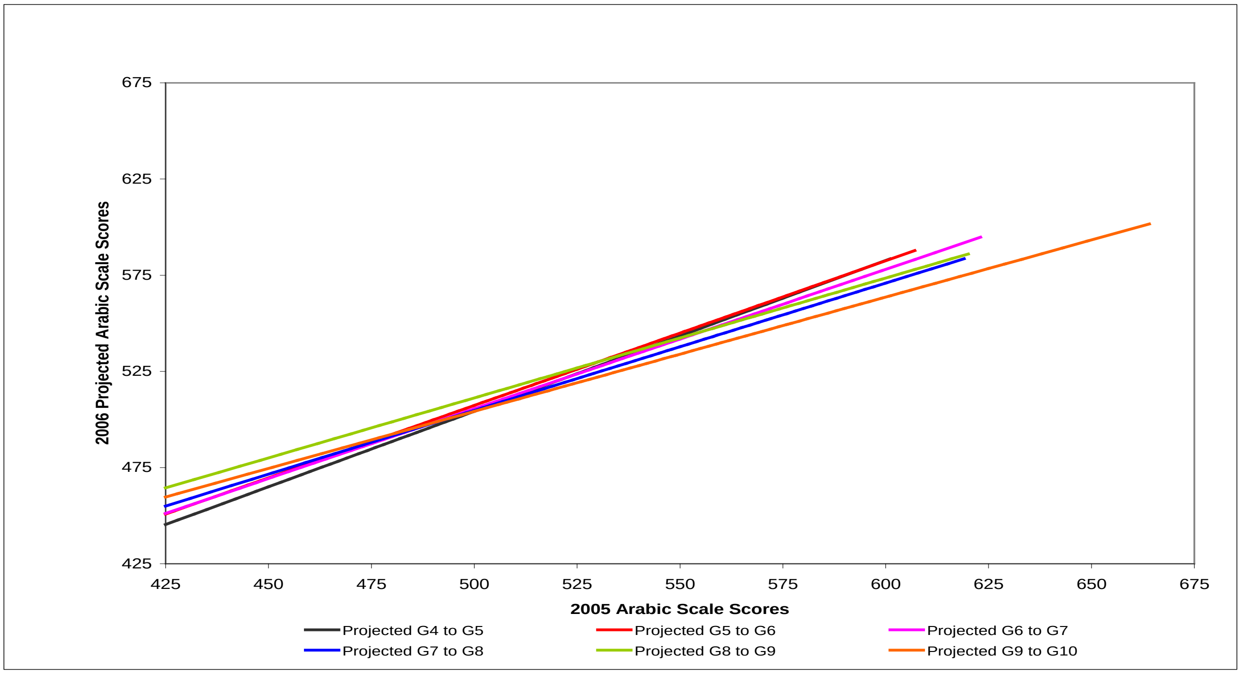

Figure 9. **Multilevel model: 2006 projected Arabic scale scores versus 2005 Arabic scale scores.**

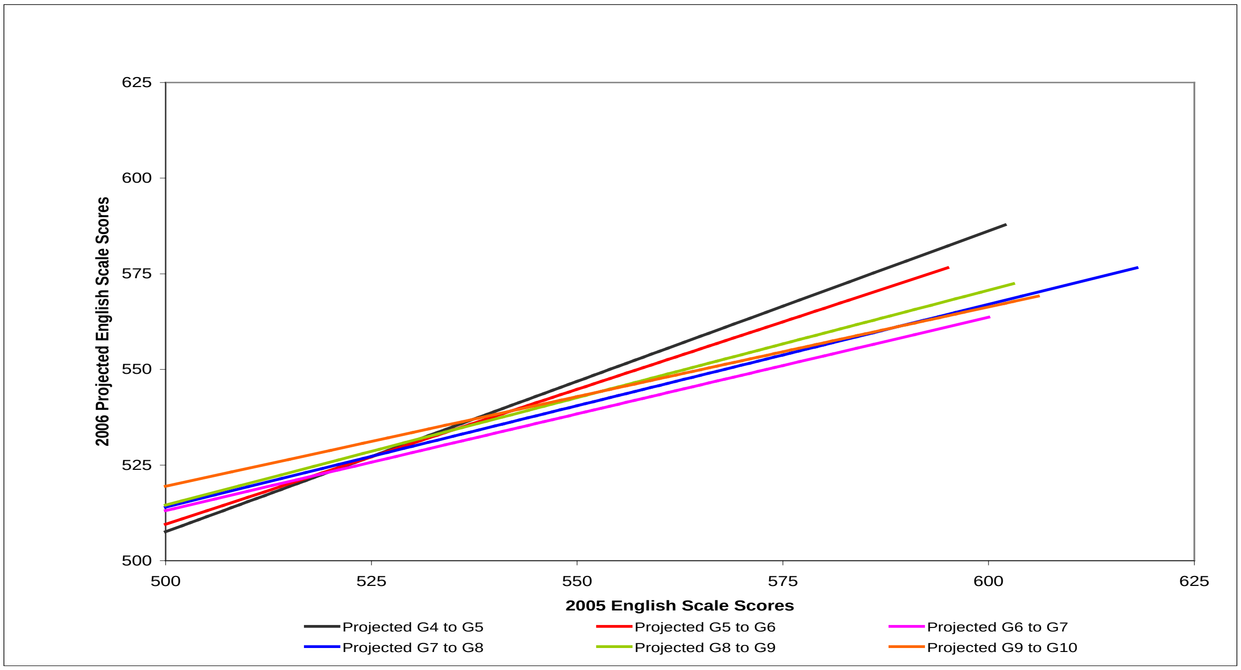

Figure 10. **Multilevel model: 2006 projected English scale scores versus 2005 English scale scores.**

As an example of using the model parameters for projecting student performance, for the student with the Arabic 2005 Grade 4 scale score of 453, the log odds of being in the Approaches Standards performance level is given by

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.6 + 0.05(453 - 467) + 0.5733 = -1.7267\,, \tag{14}$$

where -1.6 and 0.05 are the model parameter estimates from Table 11, 467 is the 2005 Grade 4 grand mean and 0.5733 is the random intercept for the school this student attended. The probability of being in the Approaches Standards performance level is then given by

$$\hat{p} = \frac{\exp(-1.7267)}{1+\exp(-1.7267)} = 0.15\,. \tag{15}$$

The projected percentages of students in the Approaches Standards performance level based on the logistic regression parameter estimates are shown in Figures 11 and 12 for Arabic and English, respectively. These projections also assume that a student will achieve the average schooling experience (school effect of 0) and therefore are based only on the fixed part of the model. Individual schools will have projection lines that deviate from the overall projection for the country. As with the empirical curves in Figures 3 and 4 and the logistic regressions in Figures 7 and 8, the multilevel projection curves show a progressive ordering across the grade levels for both content areas.

Comparisons of School Level Results Across Methods

In this section the results for the different projection methods in evaluating school performance are compared in more detail for two selected grades (Grades 6 and 8) in both content areas. The comparison is done at the school level, because the multilevel projections for each school include the school-specific random intercept effect. The empirical, OLS, and logistic projections are based on the same set of parameters estimates for all schools, while the multilevel method produces different projections by school. As an example, the projected scale scores produced by the empirical regression, OLS regression, and the multilevel linear model are presented in Figures 13 and 14 for Arabic Grade 6 and English Grade 8. For the multilevel

Table 11

Multilevel Logistic Model Results for Percentages Approaches Standards

Fixed effect	Grade 4 to Grade 5	Grade 5 to Grade 6	Grade 6 to Grade 7	Grade 7 to Grade 8	Grade 8 to Grade 9	Grade 9 to Grade 10
Arabic						
Intercept, γ_{00}	-1.6	-1.9	-2.2	-1.9	-2.4	-2.1
	(0.08)	(0.10)	(0.11)	(0.11)	(0.14)	(0.17)
Slope, γ_{01}	0.05	0.06	0.06	0.05	0.05	0.05
	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)
Variance component						
School intercept, τ_{00}	0.4	0.4	0.3	0.4	0.0	0.5
	(0.09)	(0.11)	(0.10)	(0.10)	(0.06)	(0.17)
Extradispersion	1.00	1.20	1.32	1.25	3.24	2.76
Mean λ_j	0.58	0.53	0.54	0.60	0.05	0.48
English						
Intercept, γ_{00}	-3.1	-2.9	-3.5	-4.0	-4.51	-4.4
	(0.11)	(0.11)	(0.16)	(0.19)	(0.41)	(0.22)
Slope, γ_{01}	0.12	0.11	0.09	0.11	0.12	0.0.12
	(0.004)	(0.004)	(0.004)	(0.006)	(0.014)	(0.006)
Variance component						
School intercept, τ_{00}	0.8	0.5	0.8	0.4	0.0	0.7
	(0.16)	(0.13)	(0.21)	(0.14)	(0.00)	(0.23)
Extradispersion	0.67	0.89	0.84	1.30	5.49	0.83
Mean λ_j	0.57	0.51	0.62	0.38	0.00	0.52

Note. Standard errors of the intercepts and slopes appear in parentheses under those values.

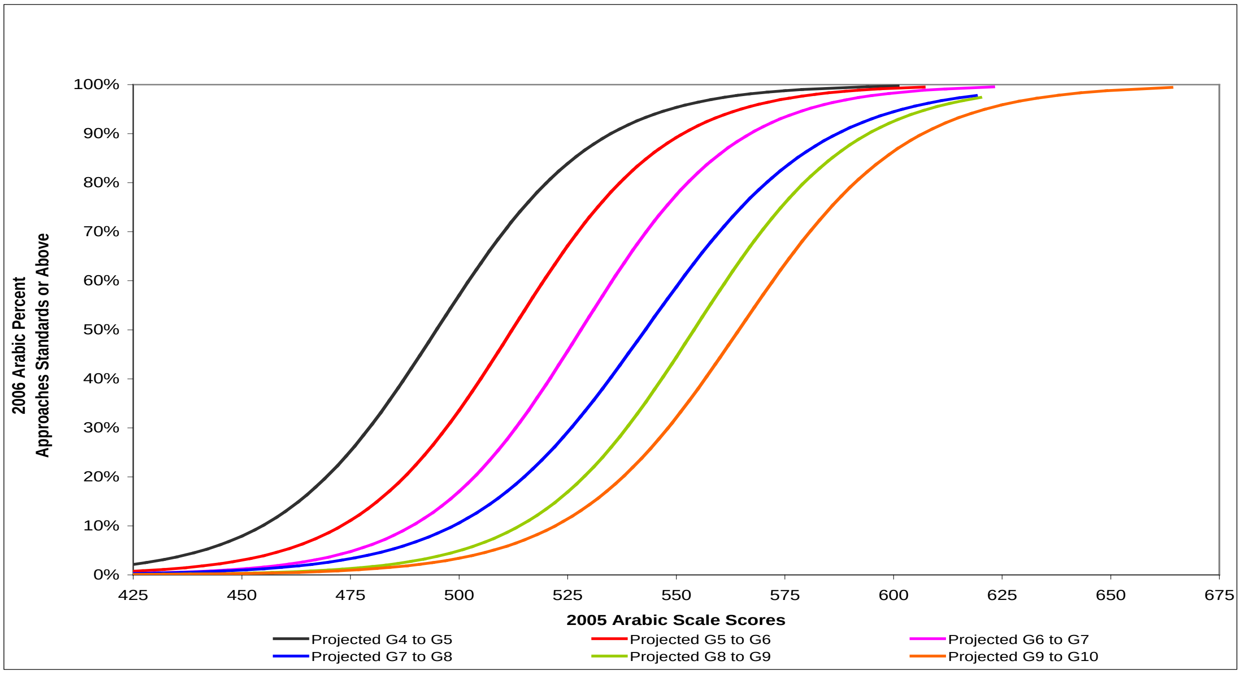

Figure 11. **Multilevel: 2006 projected Arabic percentage Approaches Standards versus 2005 Arabic scale scores.**

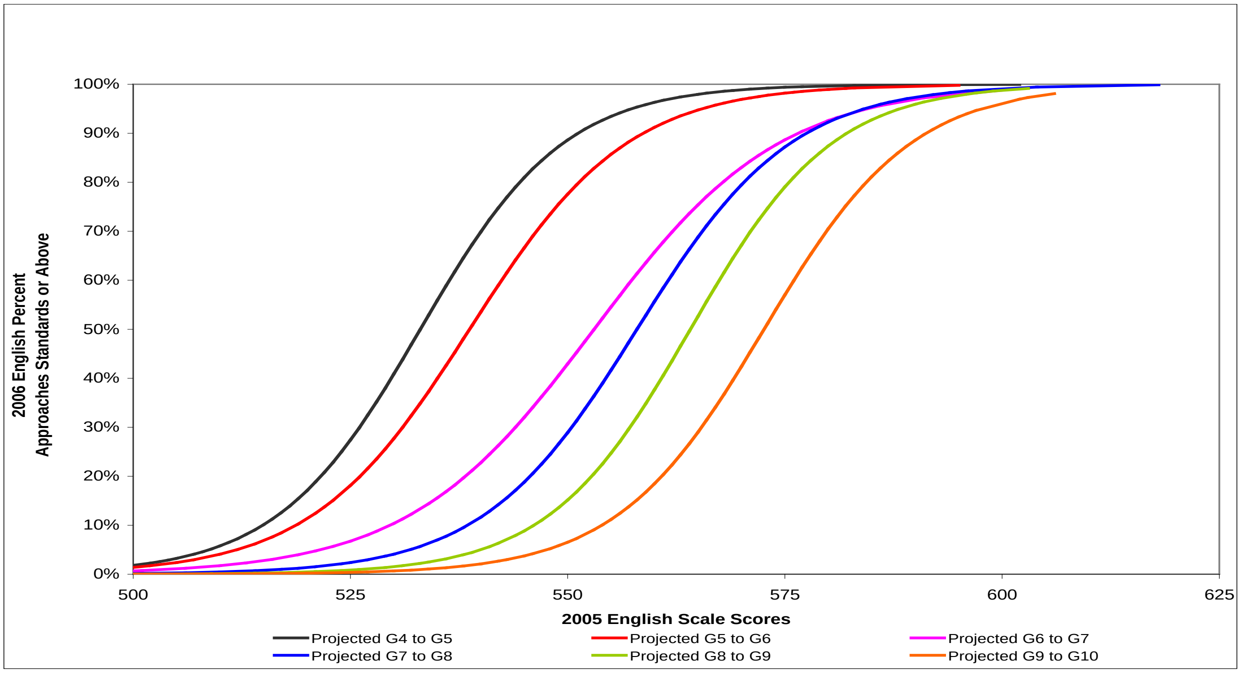

Figure 12. **Multilevel: 2006 projected English percentage Approaches Standards versus 2005 English scale scores.**

projections there are three exemplar projections lines: the Average Multilevel projection line represents the average population projection, the Low Multilevel projection line represents the projection for a low-performing school, and the High Multilevel projection line represents the projection line for a high-performing school; these three lines vary in terms of their school intercepts, u_{0j}. For Arabic, the OLS and average multilevel projections are very similar, with the multilevel procedure showing a slightly greater regression effect (i.e., flatter overall slope). For Arabic, the empirical regressions can be seen to be somewhat more curvilinear than the OLS and multilevel results. Nonetheless, the OLS and multilevel models provide an adequate fit to the data, particularly at the lower end of the scale where most of the students are scoring. For English, differences between the methods are much more pronounced, particularly in Grade 8. The disparities in the projections for English occur because the model-based OLS and multilevel estimates do not adequately fit the data, particularly among the higher scoring students.

Figure 15 displays for Grade 6 Arabic the projections for the percentage of students who are in the Approaches Standards performance category, given their scale scores in Grades 5. The results for the English projections are shown in Figure 16. The low multilevel and high multilevel projections are for the same schools shown in the multilevel projections in Figures 13 and 14. For Arabic, the plots show that the logistic regression and the average multilevel logistic projections were almost identical. Both were similar to the empirical projections, although they tend to differ at the upper end of the scale, where there are few students. In comparison to Arabic, there was more variability in the projections for Grade 8 English, reflecting the less-than-adequate fit of the multilevel logistic model. The multilevel reliability estimate for English Grade 8 is 0.38 (Table 11) reflecting that there is considerable shrinkage in the empirical Bayes estimates for the school effects, and the low and high school multilevel regressions were therefore close to the average over all schools. Also, for English Grade 8 the overall slope for the multilevel fitted line in Figure 16 was flatter than the empirical regression.

It is apparent from the figures above that the regression-based methods differ to various degrees across the range of score scales. A question of interest is how the methods compare in projecting growth for individual schools. To this end, a subset of study schools was identified that had results for at least 15 students (i.e., $n \geq 15$ matched cases for the grade pair being analyzed). For Arabic, there were 73 such schools for projecting Grade 6 and 54 schools for

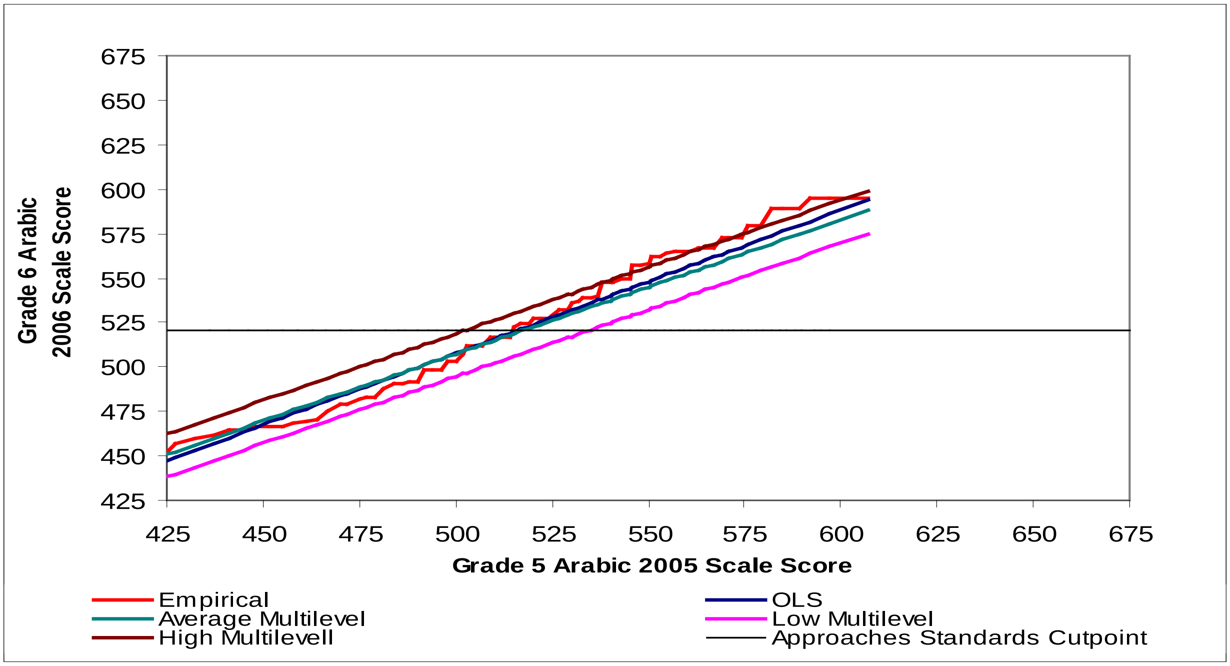

Figure 13. **Grade 6 Arabic 2006 projected scale scores by method.**

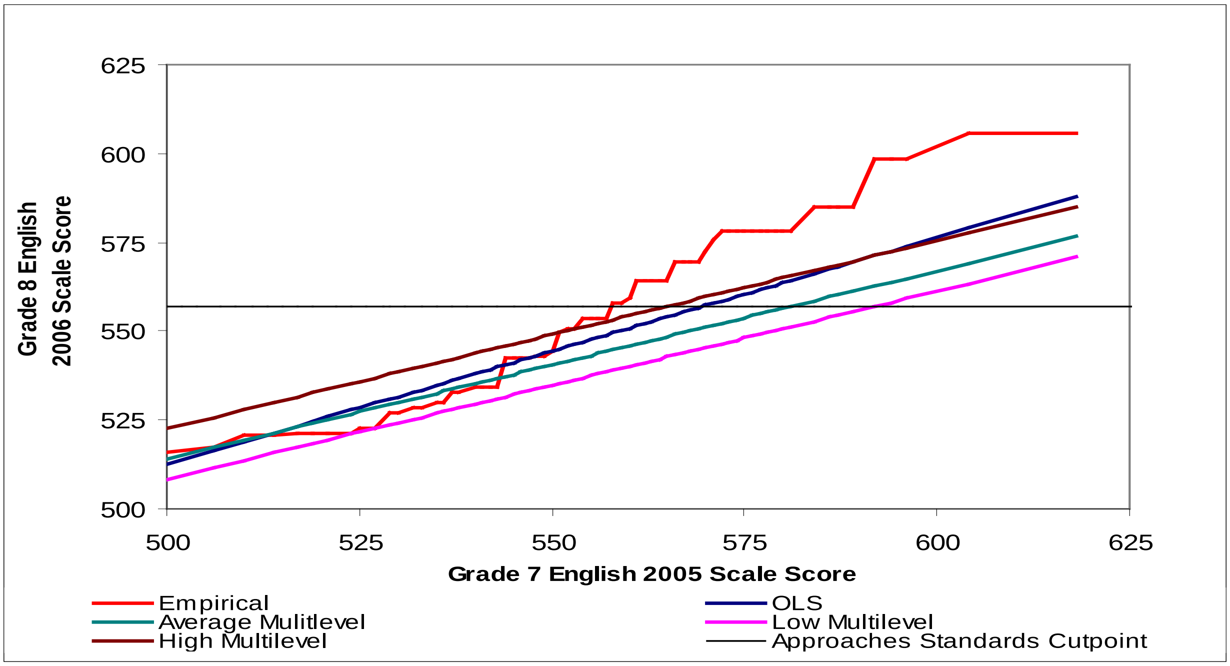

Figure 14. **Grade 8 English 2006 projected scale scores by method.**

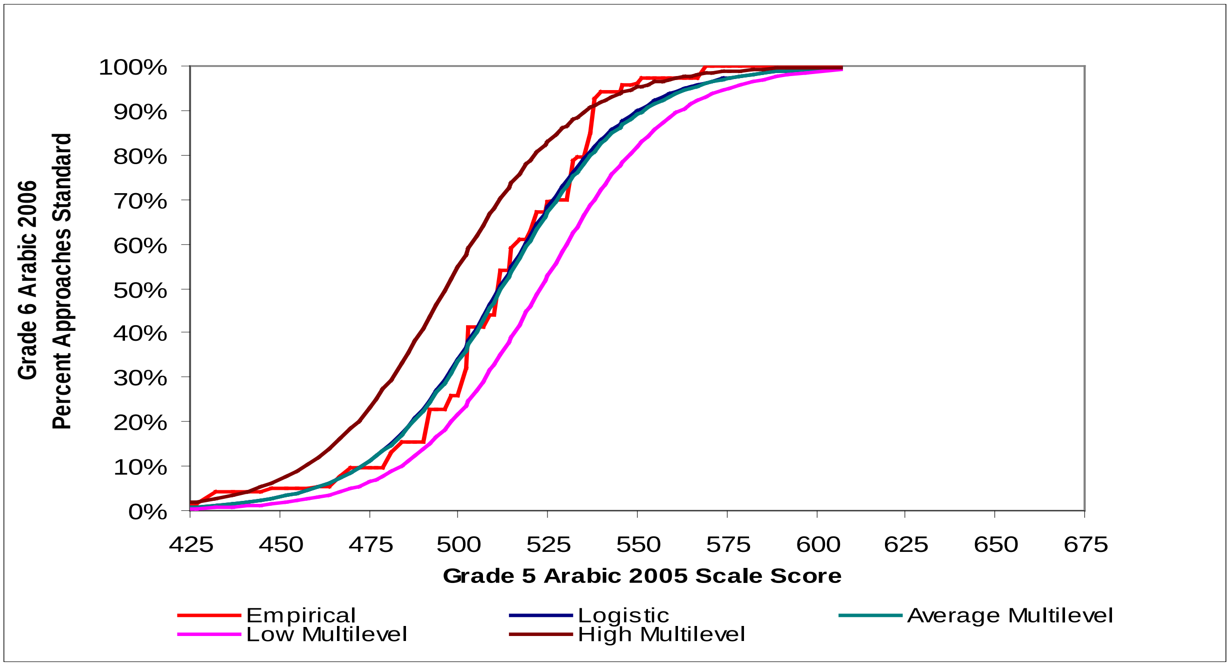

Figure 15. **Grade 6 Arabic 2006 projected percentage Approaches Standards by method.**

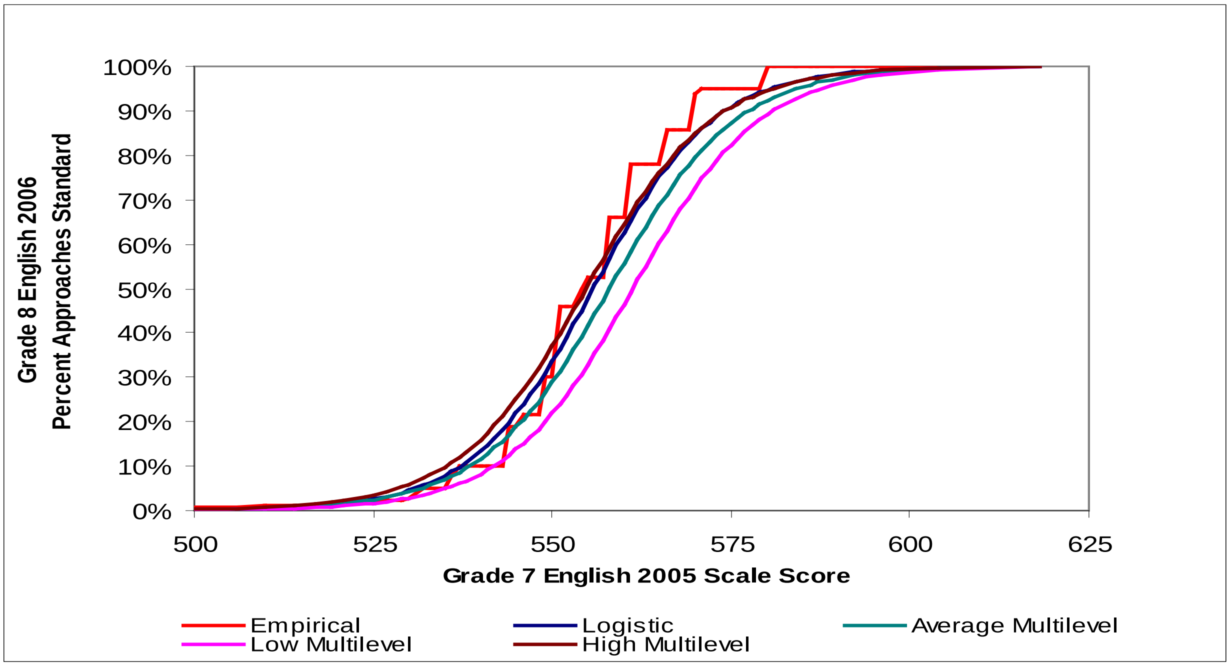

Figure 16. **Grade 8 English 2006 projected percentage Approaches Standards by method.**

Grade 8; for English there were 76 schools for Grade 6 and 49 schools for Grade 8. Figures 17 to 24 display predicted versus observed school means for these schools. The empirical and OLS or logistic predictions produce very similar results at the school level and generally show overprediction of low-scoring schools and underprediction of high-scoring schools. Another way to say this is that the empirical and OLS regressions were based on pooling together data from all students, regardless of the school they attended. When those regressions were then applied, the students in low-performing schools tended to do less well than predicted and the students in the high-performing schools tended to do better than predicted.

As expected, the multilevel models, which are focused on fitting school-level results, provide much more accurate estimates of school performance. The multilevel estimates did not show the bias trend seen for the empirical and OLS models, but did display slight bias downwards for most of the score scale range; this effect was greater in English than Arabic.

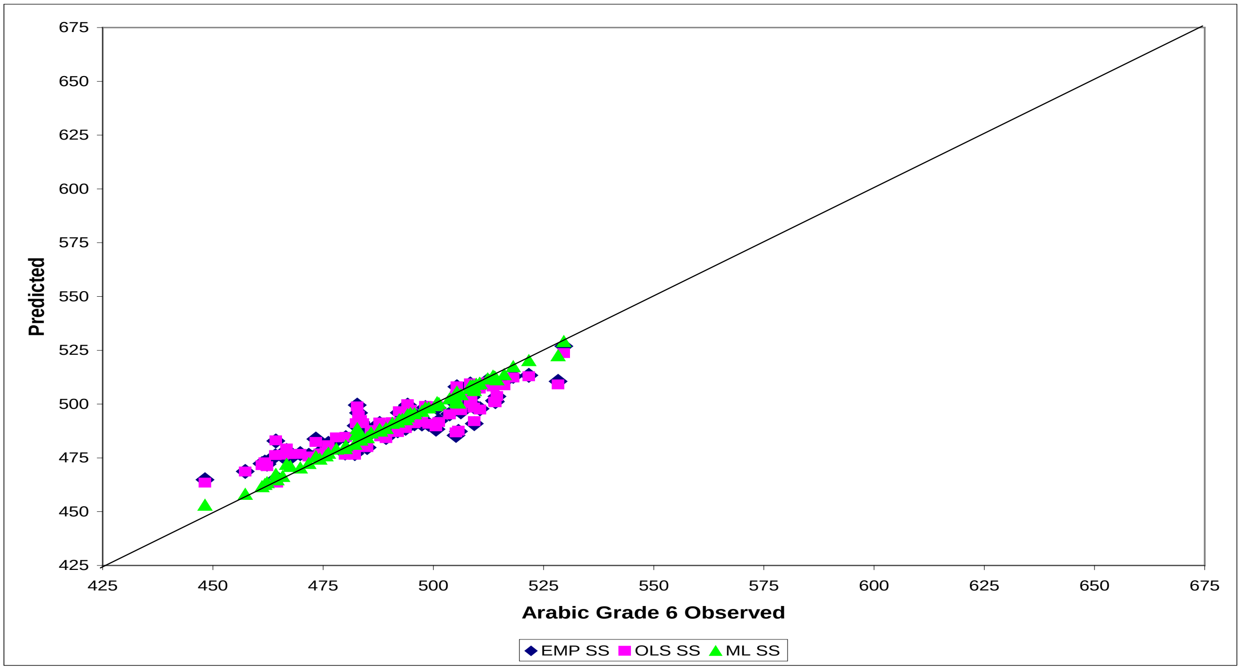

Figure 17. **Observed versus predicted school means: Arabic Grade 6 scale scores.**

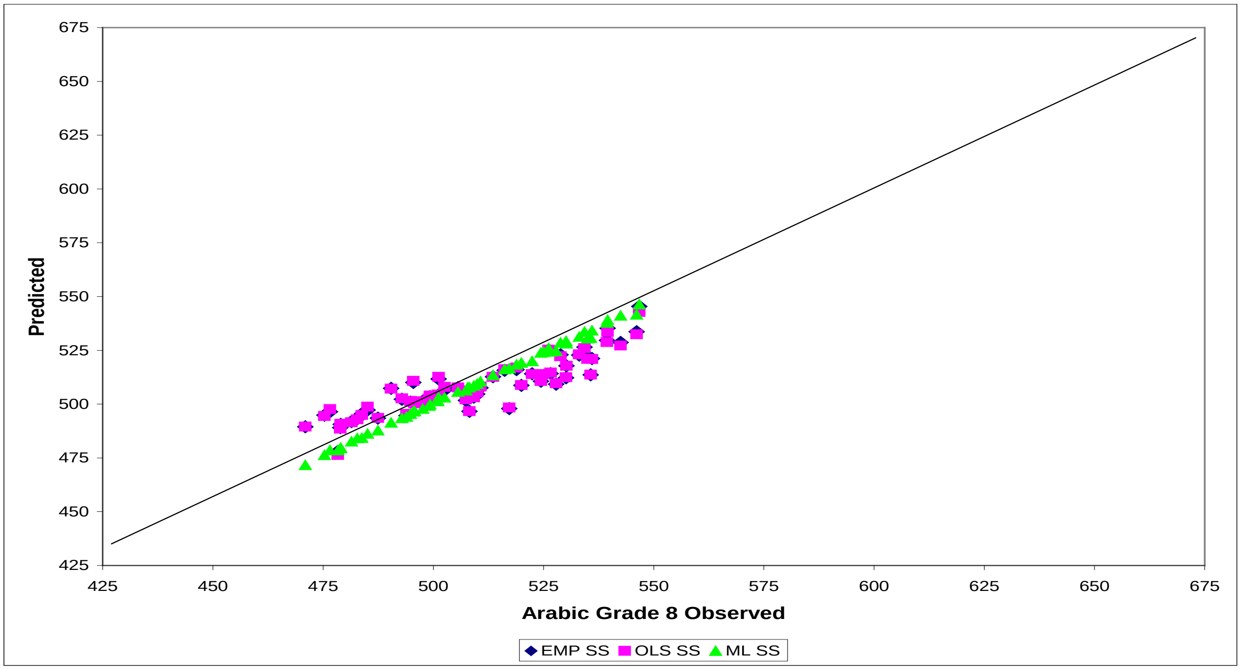

Figure 18. **Observed versus predicted school means: Arabic Grade 8 scale scores.**

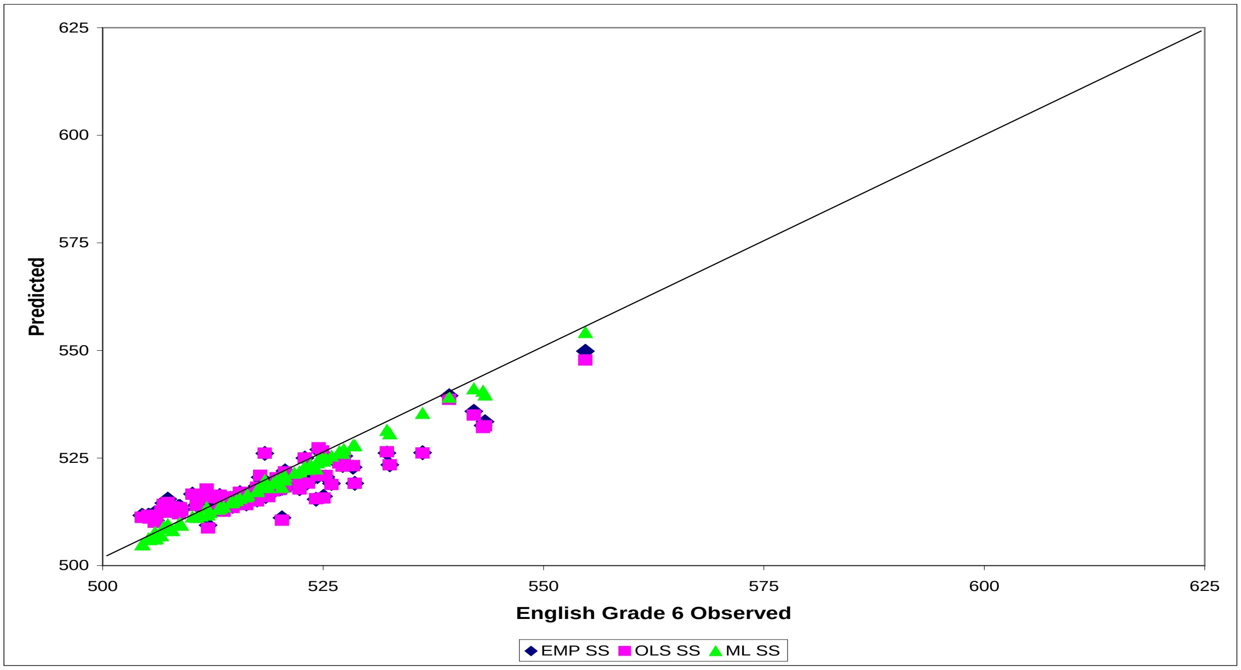

Figure 19. **Observed versus predicted school means: English Grade 6 scale scores.**

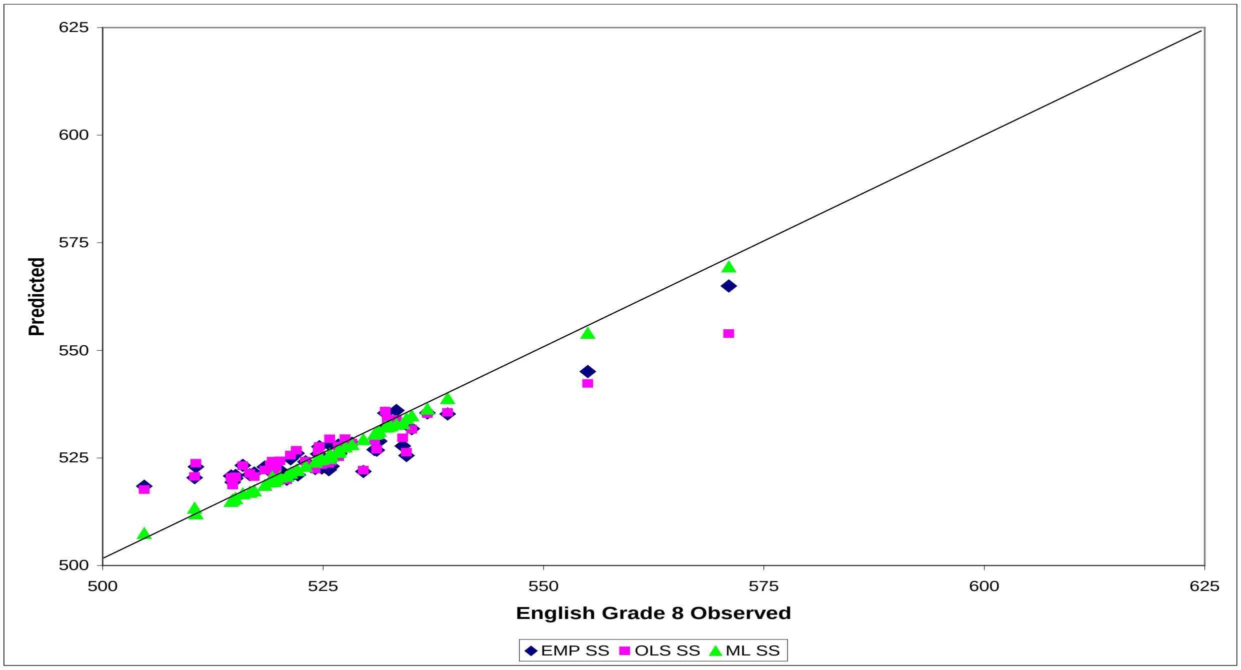

Figure 20. **Observed versus predicted school means: English Grade 8 scale scores.**

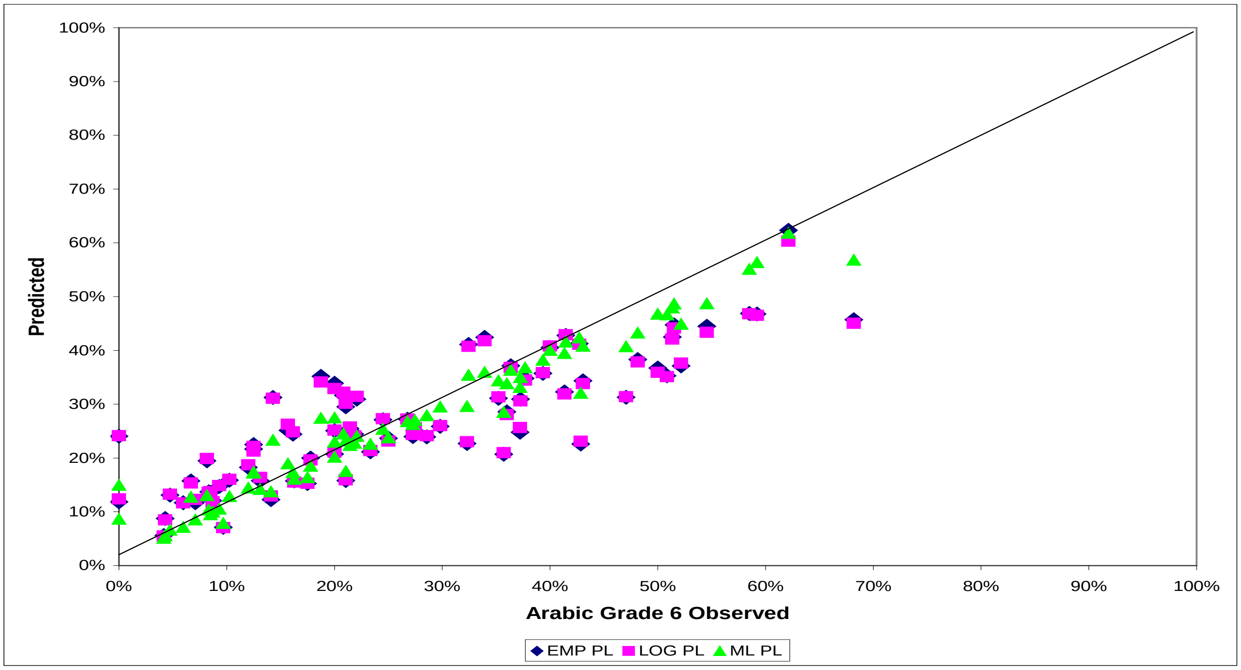

Figure 21. **Observed vs. predicted school percentage Approaches Standards: Arabic Grade 6.**

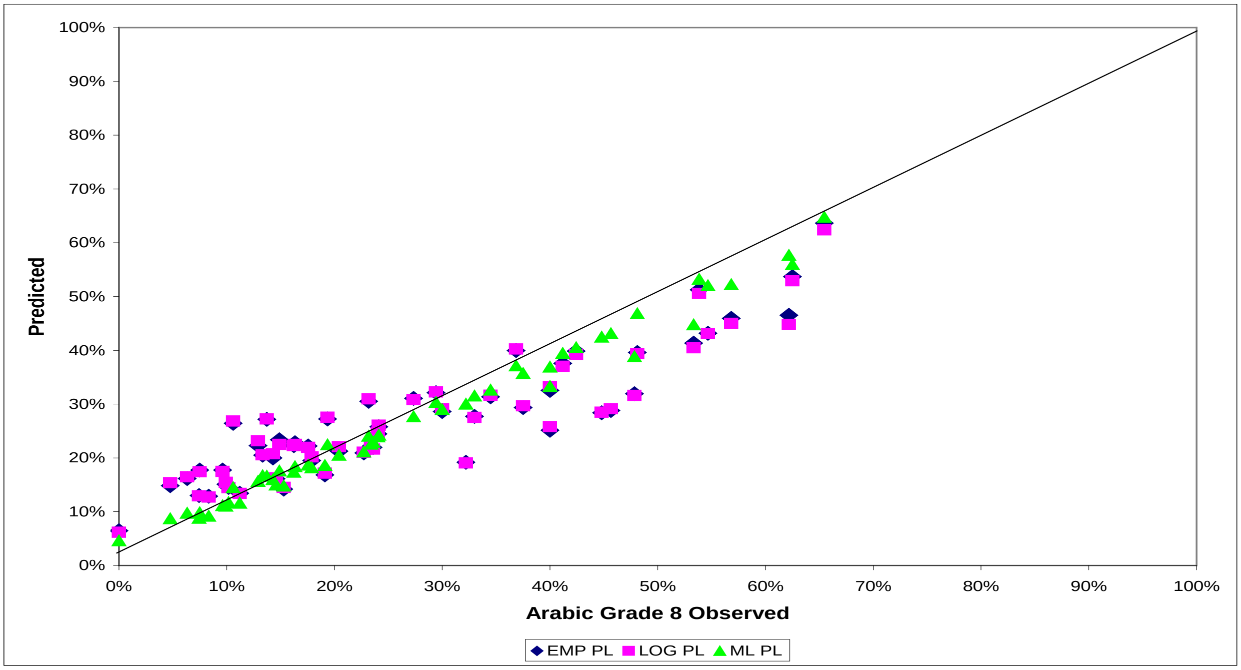

Figure 22. **Observed versus predicted school percentage Approaches Standards: Arabic Grade 8.**

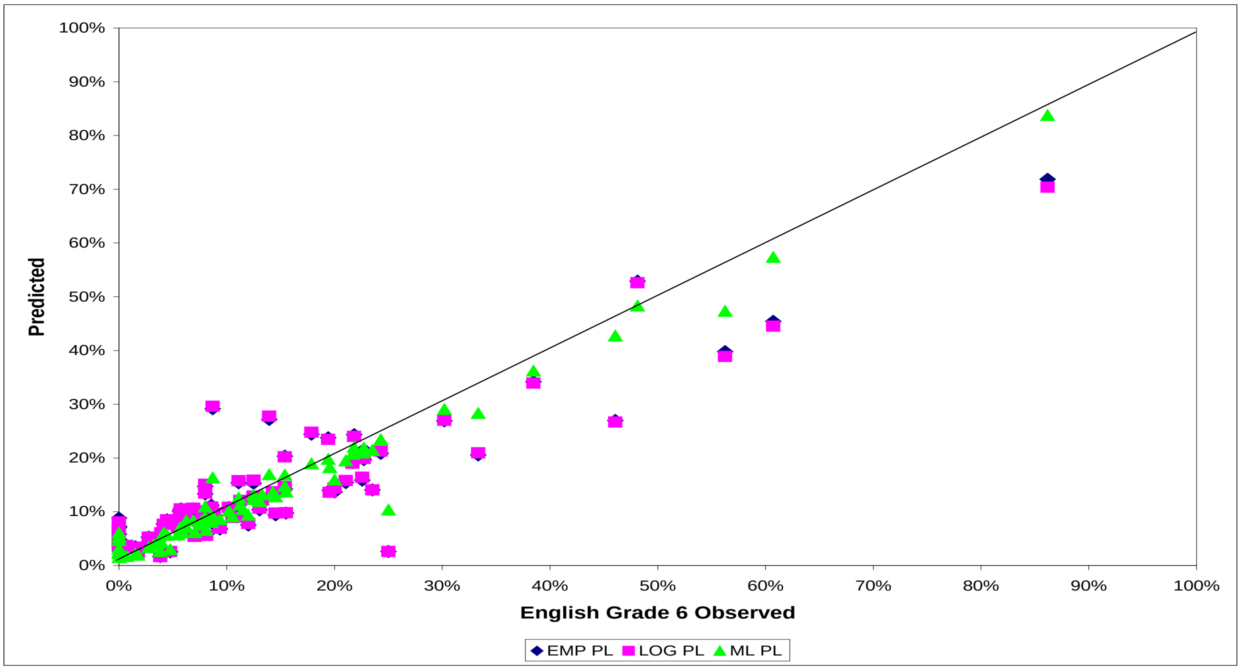

Figure 23. **Observed versus predicted school percentage Approaches Standards: English Grade 6.**

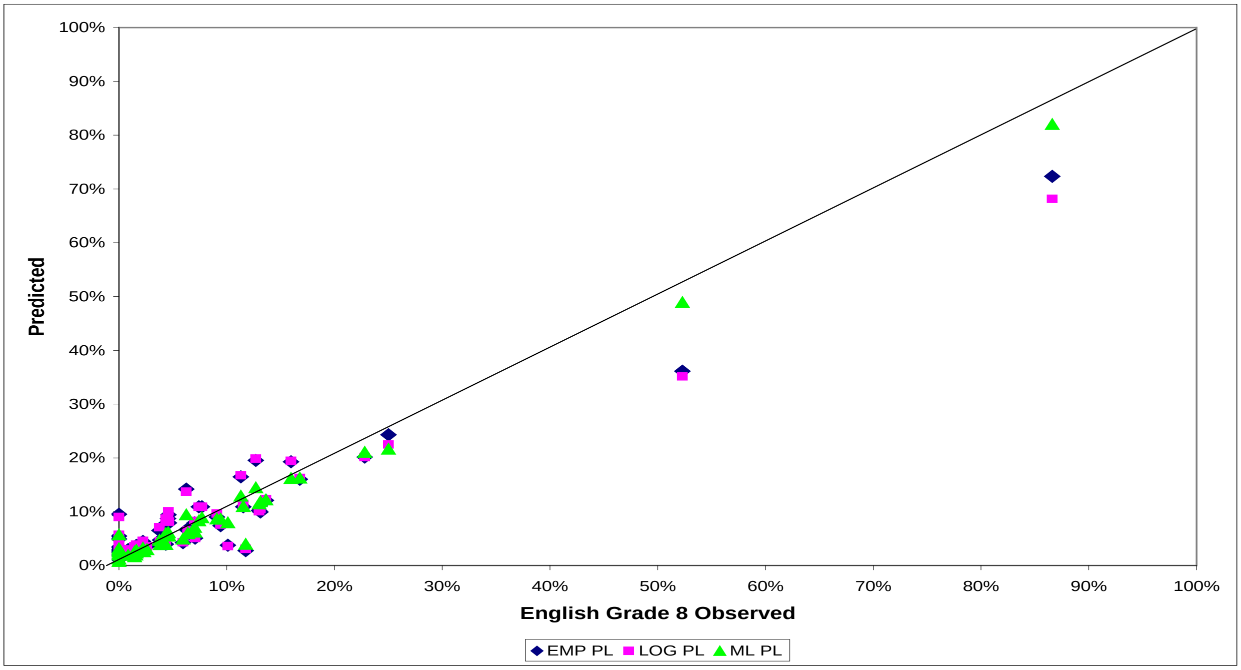

Figure 24. **Observed versus predicted school percentage Approaches Standards: English Grade 8.**

To further compare the projection methods for evaluating growth, the scale scores and percentage Approaches Standards were calculated for the same set of schools represented in Figures 17 to 24.

Scale scores. The scale scores were as follows:

- Vertical scale: Mean 2006 observed scale score minus mean 2005 observed scale score for the same students
- Empirical regression: Mean 2006 observed scale score minus mean 2006 projected scale score based on the empirical regression applied to the same students' 2005 scale scores
- OLS regression: Mean 2006 observed scale score minus mean 2006 projected scale score based on the OLS regression applied to the same students' 2005 scale scores
- Multilevel regression:
    - School intercept: The value of the school random intercept effect, u_{0j}

  - Residual: Mean 2006 observed scale score minus mean 2006 projected scale score based on the multilevel regression applied to the same students’ 2005 scale scores

Percentage reaching Approaches Standards. The percentage reaching Approaches Standards were as follows:

- Vertical scale: 2006 observed percentage Approaches Standards minus 2005 observed percentage Approaches Standards for the same students
- Empirical regression: 2006 observed percentage Approaches Standards minus 2006 projected percentage Approaches Standards based on the empirical regression applied to the same students’ 2005 scale scores
- OLS regression: 2006 observed percentage Approaches Standards minus 2006 projected percentage Approaches Standards based on the OLS regression applied to the same students’ 2005 scale scores
- Multilevel regression:
  - School intercept: The value of the school random intercept effect, u_{0j}
  - Residual: 2006 observed percentage Approaches Standards minus 2006 projected percentage Approaches Standards based on the multilevel regression applied to the same students’ 2005 scale scores

Note that the growth measures calculated for the empirical, OLS, and logistic regressions were equivalent to determining how a school’s observed growth compared to its projected growth based on applying the country-level regressions to the students in that school. In other words, the regressions accounted for the prior year’s test scores and based projected growth on the amount of growth seen by students across the country who started in the same place. For the multilevel model, the more a school exceeded the typical performance seen in general for students who started in the same place, the greater the school-specific increment to the Level I intercept. The vertical scale comparisons did not take into account the students’ starting places, but gave schools equal credit for every unit of scale score growth, regardless of how much growth was typically seen for similar students.

For the multilevel regressions, the model-based projection took into account school-specific effects, and the growth measure could be taken as the school random intercept effect, u_{0j}, which is what is typically done in value-added modeling. We also examined the residual for the multilevel model.

Means, standard deviations, and correlations between the school growth measures are contained in Tables 12 and 13. As expected, the multilevel projections for the schools closely reflected observed results for schools: the average deviation (residual) between the observed school means and the multilevel model projected means was less than one half of a scale score point. The empirical and OLS regressions showed good results at the grand mean level—with the average difference between observed and projected school means close to 1 scale score point or less—but usually greater variance in school mean deviations than the multilevel projections. (See also Figures 17 to 20.) This result makes sense: systematic deviations in the school mean are incorporated into the multilevel model parameter, u_{0j}, but are included as part of the growth measure (residual) for the empirical and OLS regressions. The means for the vertical scale simply reflect the mean school difference in scale scores from one grade to the next.

For the percentage Approaches Standards (Table 13), all three regression techniques produced overall mean growth very similar to observed growth. However, the variation between observed school means and projected ones (the residuals) was greater for the empirical regression and the logistic regression than it was for the multilevel regression (see also Figures 21 to 24). For the vertical scale the mean deviations were close to 0, reflecting the vertical moderation of the performance standards and the fact that that schools tended to perform similarly across grades.

The correlation results in Tables 12 and 13 show that for both scale scores and performance level analyses, the empirical, OLS, and multilevel school intercepts produced very similar school results, with correlations ranging from 0.90 to 1.00. The multilevel model residuals also had substantial correlations with the empirical and OLS residuals, with correlations ranging from 0.81 to 0.95. It is of further interest that the correlations between the multilevel school intercepts and the multilevel residuals were substantial (0.77 to 0.81). While the multilevel residuals had very small absolute values, there was a tendency for higher performing schools to perform a bit better than the model predicted and for lower performing schools to perform a bit worse than predicted.

Table 12

Means, Standard Deviations, and Correlations Between School Mean Growth Measures Based on Scale Score Projections

				Multilevel model	
	Vertical scale	Emp. regression	OLS regression	Residual	u_{0j}
Arabic Grade 6 (*N* = 73)					
Mean (SD)	12.4 (7.8)	0.8 (8.7)	0.8 (8.8)	-0.0 (2.0)	0.0 (7.5)
Correlations					
Vertical scale	1.00				
Emp. regression	0.94	1.00			
OLS regression	0.94	0.99	1.00		
Multilevel residual	0.84	0.88	0.88	1.00	
Multilevel u_{0j}	0.88	0.98	0.99	0.81	1.00
Arabic Grade 8 (*N* = 54)					
Mean (SD)	4.0 (9.4)	1.2 (10.9)	1.2 (11.1)	0.4 (1.5)	0.8 (10.5)
Correlations					
Vertical scale	1.00				
Emp. regression	0.90	1.00			
OLS regression	0.89	1.00	1.00		
Multilevel residual	0.68	0.81	0.82	1.00	
Multilevel u_{0j}	0.85	0.99	0.99	0.77	1.00
English Grade 6 (*N* = 76)					
Mean (SD)	6.3 (4.5)	0.3 (5.0)	0.3 (5.0)	0.0 (1.0)	0.3 (4.7)
Correlations					
Vertical scale	1.00				
Emp. regression	0.95	1.00			
OLS regression	0.95	1.00	1.00		
Multilevel residual	0.82	0.87	0.88	1.00	
Multilevel u_{0j}	0.88	0.97	0.98	0.82	1.00
English Grade 8 (*N* = 49)					
Mean (SD)	3.4 (4.2)	-1.0 (4.9)	-1.0 (5.5)	-0.1 (0.8)	-0.1 (5.6)
Correlations					
Vertical scale	1.00				
Emp. regression	0.78	1.00			
OLS regression	0.74	0.94	1.00		
Multilevel residual	0.59	0.84	0.85	1.00	
Multilevel u_{0j}	0.63	0.90	0.99	0.80	1.00

Note. Emp. = empirical, OLS = ordinary least squares.

Table 13

Means, Standard Deviations, and Correlations Between School Mean Growth Measures Based on Performance Level Projections

				Multilevel model	
	Vertical scale	Emp. regression	OLS regression	Residual	u_{0j}
Arabic Grade 6 (*N* = 73)					
Mean (SD)	-0.01 (0.09)	0.00 (0.09)	0.00 (0.09)	0.00 (0.04)	-0.01 (0.48)
Correlations					
Vertical scale	1.00				
Emp. regression	0.88	1.00			
OLS regression	0.87	1.00	1.00		
Multilevel residual	0.84	0.95	0.95	1.00	
Multilevel u_{0j}	0.77	0.93	0.94	0.81	1.00
Arabic Grade 8 (*N* = 54)					
Mean (SD)	-0.02 (0.08)	0.00 (0.08)	0.00 (0.08)	0.00 (0.03)	0.01 (0.50)
Correlations					
Vertical scale	1.00				
Emp. regression	0.76	1.00			
OLS regression	0.74	1.00	1.00		
Multilevel residual	0.59	0.89	0.89	1.00	
Multilevel u_{0j}	0.74	0.96	0.96	0.77	1.00
English Grade 6 (*N* = 76)					
Mean (SD)	0.04 (0.07)	0.00 (0.07)	0.00 (0.07)	0.00 (0.03)	0.03 (0.53)
Correlations					
Vertical scale	1.00				
Emp. regression	0.90	1.00			
OLS regression	0.89	1.00	1.00		
Multilevel residual	0.84	0.89	0.89	1.00	
Multilevel u_{0j}	0.80	0.93	0.93	0.80	1.00
English Grade 8 (*N* = 49)					
Mean (SD)	-0.02 (0.05)	-0.01 (0.05)	-0.01 (0.05)	0.00 (0.02)	-0.01 (0.42)
Correlations					
Vertical scale	1.00				
Emp. regression	0.57	1.00			
OLS regression	0.49	0.99	1.00		
Multilevel residual	0.47	0.87	0.85	1.00	
Multilevel u_{0j}	0.45	0.91	0.91	0.78	1.00

Note. Emp. = empirical. OLS = ordinary least squares.

The trends in the correlations of the vertical scale were quite similar for Arabic Grades 6 and 8 and English Grade 6. The correlations of the vertical scale growth measure and the empirical and OLS regressions were substantial in the scale score metric (ranging from 0.89 to 0.95) but lower in the performance level metric (0.74 to 0.90). Similarly, while the vertical scale tended to have substantial correlations with the multilevel school effects variable when examining scale scores (0. 85 to 0.88), these correlations were lower in the performance level metric (0.74 to 0.80). The correlations of the vertical scale with the multilevel residuals tended to be lower than the correlations with the multilevel school effects in the metric of scale scores (0.68 to 0.84) and performance levels (0.59 to 0.84). For English Grade 8, the trends for the vertical scale were similar to those just described for the other grades and content areas, but the English Grade 8 correlations all tended to be substantially lower (0.63 to 0.78 for scale scores, 0.45 to 0.57 for performance levels).

The correlation results provided in Tables 12 and 13 are shown in scatterplots for Grade 6 Arabic and Grade 8 English projections in Figures 25 to 28. (The multilevel school intercepts are not included in these plots.) Using Qatar's system of classifying schools based on school total enrollment across all grades, schools were classified into small (*n* = 15–200), medium (*n* = 201–500) and large (*N* > 500).

A school mean is affected by performance throughout the distribution of scores, but the percentage of students reaching a performance level is affected by performance of students near that performance level. How much difference did it make to use scale score results to measure school growth versus using percentages of students reaching Approaches Standards? Table 14 displays the school-level correlations between results in these two metrics within each projection method. For Arabic Grades 6 and 8 and for English Grade 6, the use of school means or percentage reaching Approaches Standards produced correlations ranging from 0.73 to 0.87; for English Grade 8 the correlations were substantially lower, ranging from 0.49 to 0.72. These results are consistent with the fact that in English—especially Grade 8 English—there were small percentages of students reaching the Approaches Standards performance level, so evaluating school growth in terms of scale score means produced different conclusions from evaluating school growth in terms of percentages reaching the challenging target performance level.

Table 14

Correlations Between Growth Based on Scale Scores Projections and Growth Based on Performance Level Projections Within the Same Projection Method

	Correlation
Arabic Grade 6 (*N* = 73)	
Vertical scale	0.78
Empirical regression	0.81
OLS/Logistic regression	0.82
Multilevel residual	0.85
Multilevel u_{0j}	0.87
Arabic Grade 8 (*N* = 54)	
Vertical scale	0.78
Empirical regression	0.84
OLS/Logistic regression	0.85
Multilevel residual	0.87
Multilevel u_{0j}	0.87
English Grade 6 (*N* = 76)	
Vertical scale	0.73
Empirical regression	0.76
OLS/Logistic regression	0.77
Multilevel residual	0.79
Multilevel u_{0j}	0.82
English Grade 8 (*N* = 49)	
Vertical scale	0.49
Empirical regression	0.52
OLS/Logistic regression	0.67
Multilevel residual	0.49
Multilevel u_{0j}	0.72

Note. OLS = ordinary least squares.

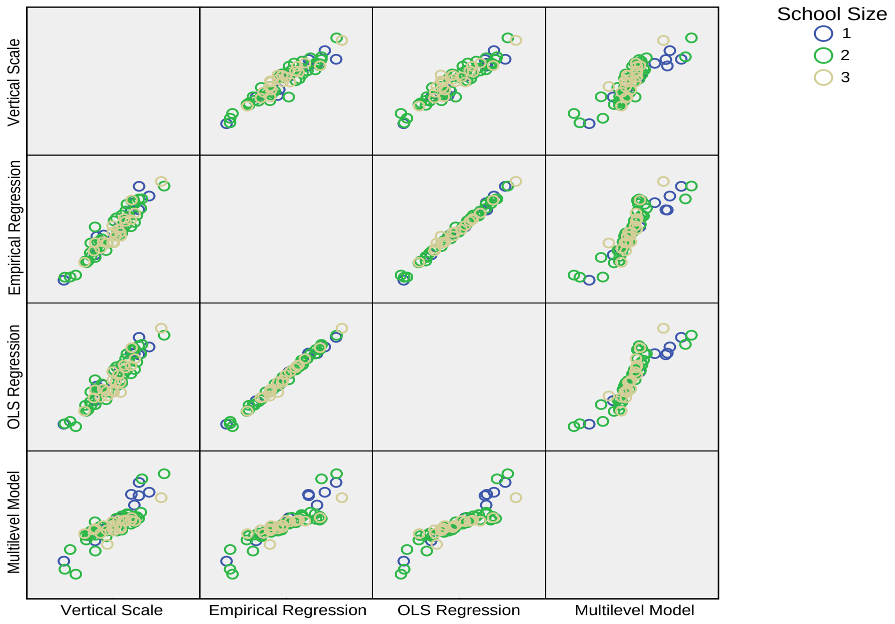

Figure 25. **Bivariate plots of school mean difference between Arabic Grade 6 observed 2006 scale score and projected 2006 scale score.**

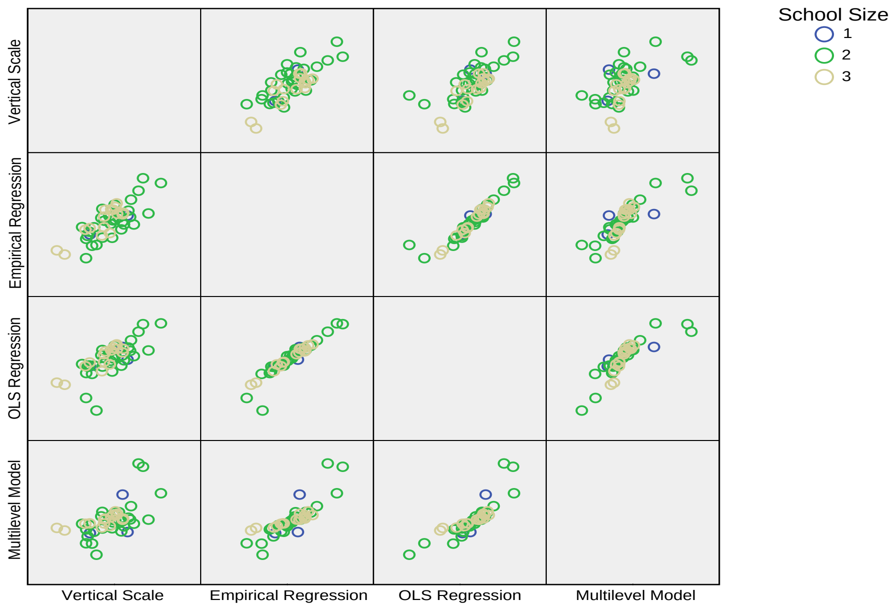

Figure 26. **Bivariate plots of school mean difference between English Grade 8 observed 2006 scale score and projected 2006 scale score.**

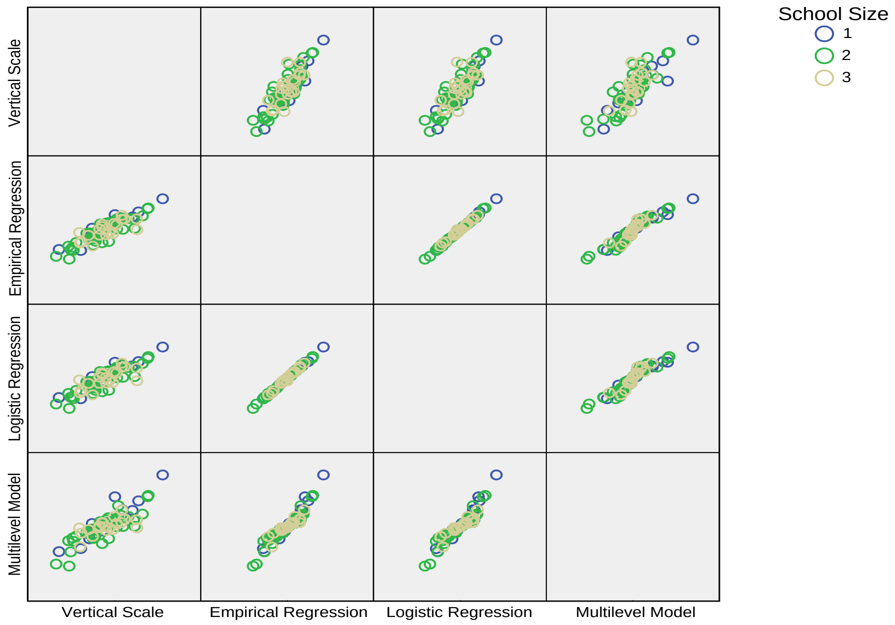

Figure 27. **Bivariate plots of school mean difference between Arabic Grade 6 observed 2006 percentage Approaches Standards and projected 2006 percentage Approaches Standards.**

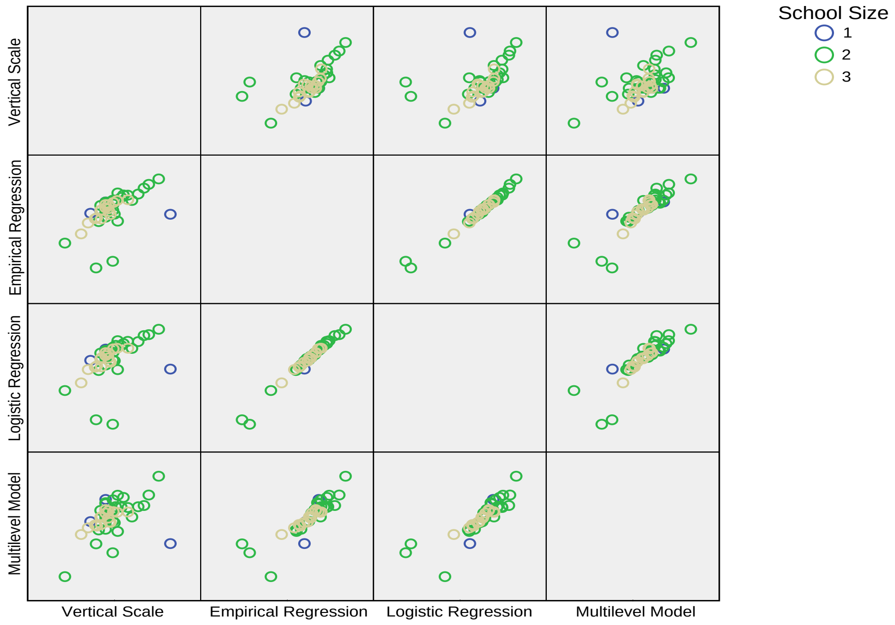

Figure 28. **Bivariate plots of school mean difference between English Grade 8 observed 2006 percentage Approaches Standards and projected 2006 percentage Approaches Standards.**

Summary and Discussion

There is interest in measuring longitudinal growth using educational assessments. While vertical scales offer one method for doing so, vertical scales often are not available or they have inherent limitations. This paper examined alternatives to vertical scales for measuring longitudinal growth and for producing school-level growth measures. The alternatives examined were the following:

- Empirical cross-grade regressions
  - Regression of scale score on the students’ scale scores from the previous grade
  - Regression of the dichotomous variable Below Standards/Approaches Standards on the students’ scale scores from the previous grade
- OLS and logistic regressions
  - Linear regression of scale score on the students’ scale scores from the previous grade
  - Logistic regression of the dichotomous variable Below Standards/Approaches Standards on the students’ scale scores from the previous grade
- Multilevel analyses
  - Multilevel linear regression of scale score on the students’ scale scores from the previous grade, taking school context into account by modeling the clustering of students nested within schools
  - Multilevel logistic regression of the dichotomous variable Below Standards/Approaches Standards on the students’ scale scores from the previous grade, taking school context into account by modeling the clustering of students nested within schools

The student data used for the comparisons were English and Arabic Grades 4 to 10 in Qatar. An advantage of the use of these data was that operational vertical scales were in place for this assessment program. Thus, the growth measures that did not require a vertical scale could be compared with the results produced by the vertical scale.

Applicability of Each Method

These assessments reflect challenging Qatari content and performance standards, and the tests were difficult for the students, especially in English. Mean p-values for the tests ranged from 0.42 to 0.51 for Arabic and from 0.29 to 0.41 for English. Reliabilities ranged from 0.85 to 0.89 for Arabic and from 0.76 to 0.90 for English. Thus, assessments had acceptable reliabilities, but the generality of these results was affected by the test difficulty.

The percentages of students whose scores could be matched from one grade to the next ranged from 62% to 80%, with the lower grades having higher percentages matches than the upper grades. Cross-grade correlations of student scale scores ranged from 0.68 to 0.74 for Arabic and from 0.59 to 0.73 for English. The correlations between students' scale scores in 2005 and their proficiency levels in 2006 ranged from 0.54 to 0.62 in Arabic and from 0.51 to 0.67 in English. While higher match rates would be desirable, the match rates and the cross-grade correlations were acceptable for this study.

Vertical scales and performance levels. In the vertical scaling that was conducted using IRT as part of the operational testing program, procedures and results appeared sound (ETS, 2006). Appropriate practices were followed in terms of selection and administration of cross-grade linking items, and model fit was good. The results conformed to expectations of vertical scaling in terms of consistency of item parameter estimates across grades, progression of test characteristic functions across grades (demonstrating increasing test difficulty as grade increased), and progressions of scale score distributions (e.g., higher scale score means as grade increased). Thus, based on the procedures followed and the results, the Arabic and English vertical scales would be accepted as sound vertical scales.

The performance levels were also developed using widely used and accepted procedures (ETS, 2006). The procedure for setting performance standards produced, by design, a progression of standards with similar impact (i.e., percentages of student reaching the standards) by grade. One difference between the Qatari standards and performance standards set in many states within the United States is that the Qatari performance standards were generally more challenging, when measured in terms of impact. The percentages of students at or above the Approaching Standards performance level ranged from 23% to 27% for Arabic and from 9% to 14% for English.

Thus, with the general caution about the difficulty of the assessments and performance levels, the vertical scales and performance levels would be considered acceptable for use and provide a reasonable benchmark for comparing methods of measuring growth.

Empirical regressions. Empirical regressions have no assumptions, per se, at the student level, but the usability of their results in an applied setting is limited by their expected stability and reasonableness. In this setting, the cross-grade correlations were modest, but adequate, and the slopes of the regressions were clearly and significantly positive. In order to produce nondecreasing regressions, which is an essential requirement for usable results in this setting, only minor smoothing was required. The resulting empirical regressions did appear as step functions, but the sample sizes, especially for the lower grades, appeared adequate to produce reasonably regular results in both the scale score and performance level metrics. The major trends of the empirical regressions were linear in the scale score metric (Figures 1 and 2), with some flattening in the tails of the distributions; the performance level regressions appeared generally logistic in shape (Figures 3 and 4).

Ordinary least squares and logistic regressions. For Arabic scale scores, the linear OLS procedure produced regression lines that were similar to the empirical regressions through the lower part of the scale where most students scored, although the OLS regressions were generally flatter than the empirical ones at the higher end of the scale. (Compare Figures 1 and 5.) For English scale scores, the linear OLS lines also fit the empirical ones for the lower part of the scale for the lower grades, but above Grade 6, the OLS regressions were much flatter than the empirical ones for the higher end of the scale where there were few students. (Compare Figures 2 and 6.) For the performance levels, the logistic OLS regressions fit the empirical ones fairly well for most of the grades and content areas at the lower score levels, where most of the students were scoring, but not as well at the higher score levels. (Compare Figures 3 and 7 with Figures 4 and 8.)

Multilevel models. The multilevel models take into account the clustering of students that are nested within schools and thereby model school effects on student performance; the adjusted intraclass correlations, which reflect the remaining proportion of variance due to clustering within schools after accounting for 2005 scale scores, ranged from 0.10 to 0.22 for Arabic and from 0.07 to 0.21 for English. For scale score projections, the results obtained were similar to the OLS regressions. However, for projections of performance levels, the variance components and fit statistics indicate poor model-data fit for most of the Arabic and English projections due to the low variability across schools in the number of students that were in the Approaches Standards performance level.

Summary

At the student level, there was some variation in the results of the different regression-based growth measures, particularly at the higher ends of the distributions, and especially for English. If the empirical, OLS, or multilevel methods were used at the individual student level, they could produce different projections of the amount of growth expected for each student. The following section describes the differential effects of the methods when applied at the school level.

School-level results. A common request in educational assessment is for a summary measure of student academic growth that can be used as part of evaluations of school effectiveness. In this study, the four methods were applied in the scale score metric and in the performance level metric and examined in detail for Arabic and English, Grades 6 and 8. Results were compared for 49 to 73 schools, all with at least 15 students. For the vertical scale, schools were evaluated by the difference in their observed scale score means and performance levels means between 2006 and 2005 for the same students. For the regression methods, the schools were evaluated by the difference between their 2006 observed mean and the mean projected by the regression technique for the same students. For the multilevel model, the school intercepts, u_{0j}, were also examined.

For Arabic Grades 6 and 8 and English Grade 6, the three regression methods (empirical, OLS, multilevel school) produced very similar orderings of school mean growth, in both the scale score metric and the performance level metric (correlations ranging from 0.81 to 1.00. (See Tables 12 and 13.) Thus, with these data, the choice of regression method had only a modest effect on identifying the schools with the greatest and least mean growth. Correlations were lower for English Grade 8.

Considering Arabic Grades 6 and 8 and English Grade 6, the correlations of the vertical scale growth measure with the empirical and OLS regressions and the multilevel school effects were substantial in the scale score metric (ranging from 0.85 to 0.95), but lower in the performance level metric (0.74 to 0.90). The correlations of the vertical scale with the multilevel residuals tended to be lower than the correlations with the multilevel school effects in the metric of scale scores (0.68 to 0.84) and performance levels (0.59 to 0.84). For English Grade 8, the trends for the vertical scale were similar to those just described for the other grades and content areas, but the English Grade 8 correlations all tended to be substantially lower (0.63 to 0.78 for scale scores, 0.45 to 0.57 for performance levels).

Thus, within grades and content areas where performance was less extreme (Arabic Grades 6 and 8 and English Grade 6), vertical scaling produced school-level results that were quite similar to the regression techniques in the scale score metric. There are differences in the assumptions of the methods. For example, in vertical scaling, one unit of scale score growth is considered to be equivalent, regardless of whether students start out low performing or high performing. In fact, different amounts of growth are seen (at the country level) for different parts of the distribution and different grade pairs. This is seen clearly in Figures 1 and 2. However, although the methods differ at the student level, at the school level, vertical scaling and the regression techniques produced similar results for scale scores.

The empirical regressions are fairly consistent across grades but not identical. This can mean (a) that the vertical scales are not completely accurate (i.e., a 1-point change in scale scores does not mean the same thing at different grades), and/or (b) that teachers/curricula are not equally effective in different grades, and/or (c) that there are other differences (e.g., student motivation, drop outs, changes from elementary to middle school) affecting the level of growth in student scale scores between grades.

We can examine these possible explanations in more detail. It is possible that the vertical scaling is not delivering on its promise of equivalent score units throughout the scale. As described earlier, the vertical scaling was conducted following good practice and did not demonstrate any problems. Thus, no flags were raised during the vertical scaling process that provided concerns about the scaling. However, no matter how reasonable the procedures and results are from a vertical scaling, scaling tests that intentionally vary in content and difficulty cannot, by definition, produce equated scores—they can only produce linked scores. The precise equivalence of those scores will depend on the details of the content of those changing tests and how the changing content relates to what is being taught in each grade, what students learn, and how they perform on the assessment. Thus, even a perfectly developed vertical scale cannot be expected to produce score units that show uniform growth for all scale scores at all grades. It is to be expected that regressions (which take into account where students were at the end of the previous grade and the amounts of growth typically seen across the country for similar students at matching grades) will produce different results than vertical scales.

Furthermore, regression procedures display regression effects—students are not expected to be as extreme in their performance on the dependent variable as they were on the independent variable; the lower the correlation between the variables, the greater the regression effect. The student-level cross-grade scale score correlations were in the low 0.70s for almost all of the Arabic grades but for only two of the English grade-pairs. The lower correlations (0.59 to 0.61) for the predictions of the English Grades 7 to 10 produced substantial regression effects for those grades. Regression effects do not occur for the vertical scaling.

This study focused on growth between adjacent grades. In that setting, differences in regressions for different grade pairs, such as those seen in Figures 1 and 2, will not affect the results. If longitudinal data are examined over three or more years, differences in between-grade regressions can contribute to greater variation between the results of vertical scales and regression methods.

In the performance level metric, there was less similarity between the vertical scale and the other procedures. For performance levels, the vertical scaling school growth measure was the difference between the percentage of 2006 students who were at Approaches Standards minus the percentage of those same students who were at Approaches Standards in the previous grade. The performance levels were set using a combination of the modified Angoff judgment-based method at Grades 5, 8, and 11, and linear interpolation/extrapolation in the metric of percentages of students scoring at or above the cut scores at the other grades. Thus, these standards were set using appropriate measurement practice, but they did not consider amounts of change in performance levels normally seen between grades. Furthermore, relatively small percentages of students reached the Approaches Standards category (from 23% to 27% for Arabic and 9% to 14% for English). These effects, in addition to the instructional effectiveness and motivation effects described above, influence the consistency of the school performance level results.

There are many factors affecting student performance, and attempting to determine causal effects on academic performance from correlational data must be done carefully if sound conclusions are to be reached. It was noted with the present study that residuals were correlated with previous performance; that is, the models tended to underestimate the performance of students from high-scoring schools and overestimate the performance of students from low-scoring schools. Such results can be caused by correlations among factors affecting school performance, such as parental socioeconomic status and education, and/or students from similar backgrounds tending to attend the same schools. Ladd and Walsh (2002) have found that measurement error in the predictor variables can also contribute to such a result. Stuart (2007) provides a succinct description of good practice—including the use of control groups and/or propensity scores based on background variables—that can contribute to appropriate conclusions being drawn from studies of educational interventions.

Conclusions. Our conclusions were as follows:

1. Cross-grade regressions can be developed with or without the existence of a vertical scale.
2. Vertical scales and cross-grade regressions can show similar school-level results, particularly in the scale score metric. Differences between the methods appear more likely in the performance level metric than the scale score metric and for grade pairs with more extreme performance.
3. Growth measures in the performance level metric are likely to show different results from mean results in the scale score metric, particularly if the percentage of students reaching the standard is very high or low.
4. If it is desired to take into account typical amounts of growth actually observed between grades, a regression method should be used. If it is desired to treat every scale score unit, from every grade and part of the distribution, as equivalent and not take into account typical growth, a vertical scale should be used.
5. If a regression method is used, it is appropriate to use the least restrictive model that produces accurate results.
6. If growth at the individual student level is the focus of attention, or if the group for which analyses might be done is variable or unknown, then student-level regression is appropriate. However, such a model should be evaluated for possible biases in results at the group level. If analyses will only be conducted at the school level, then a multilevel model, which models school effects, is likely to be more accurate than a student-level model.

7. If sample sizes are sufficiently large, empirical regressions can be used. If empirical regression is used, its fit to results at the group level should be evaluated.

8. If sample sizes are smaller and/or a simpler model is desired, OLS regression using a linear model (in the scale score metric) or a logistic model (in the percentage reaching a standard metric) can be used. If an OLS or logistic model is used, its fit to the data should be evaluated at the student level and the school level.

9. If results indicate clustering of scores within schools, a multilevel model would be more appropriate given sufficient variability in the data and sufficient sample size at the school level. If a multilevel model is used, the fit of the model to the data should be evaluated.

10. Regression models that adequately represent or fit the observed data appear likely to produce very similar rank ordering of schools.

Cautions. The study results should be approached with the following thoughts in mind:

1. This study is based on the analysis of two years of matched data for two content areas (English and Arabic) at Grades 4 to 10 in the country of Qatar. The case counts (2,940 to 5,270) and cross-grade matching percentages (62% to 80%) were acceptable but not high. The Qatari content and performance standards are challenging for their students. The results of this study bear replication over more years and in other locales and content areas, and with assessments with a variety of difficulty levels.

2. Growth results based on large-scale assessments—no matter how they are calculated—are just one piece of information and should not be used in isolation in evaluating school effectiveness.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions.* New York: Wiley.

Bryk, A. S., & Raudenbush, S. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications, Inc.

Doran, H., & Izumi, L. (2004). *Putting education to the test: A value-added model for California*. San Francisco: Pacific Research Institute.

Drury, D., & Doran, H. (2003). *The value of value added analysis. National School Boards Association Policy Research Briefs*, 3(1), 1–4.

ETS. (2006, April). *Qatar Comprehensive Educational Assessment technical report, Arabic and English assessments Spring 2005 administration*. Princeton, NJ: Author.

ETS. (2006, January). *Qatar Comprehensive Educational Assessment (QCEA) performance level setting results from the workshops of 6-7, 11-15 December 2005 Doha, Qatar*. Princeton, NJ: Author.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8,* 41–55.

Heck, R. H., & Thomas, S. L. (2000). *An introduction of multilevel modeling techniques.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: World Book.

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review, 21,* 1–17.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society,* Series B, 34.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 17,* 159–176.

Raubertas, R. F. (2005, February 25). *Pooled adjacent violators algorithm (PAVA) implemented entirely in S.* Message posted to http://www.biostat.wustl.edu/archives/html/s-news.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (pp. 293). Thousand Oaks, CA: Sage Publications.

Roberts, J. S., & Ma, Q. L. (2006). IRT models for the assessment of change across repeated measurements. In R. W. Lissitz (Ed.), *Longitudinal and value added models of student performance* (pp.100-127). Maple Grove, MN: JAM Press.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation, 8*(3), 299–311.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210.

Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher, 36*(4), 187–198.

Yen, W. M., Lu, Y., Smith, R. L., & Patz, R. (2006). *Longitudinally comparable CST scores feasibility analyses.* Unpublished manuscript

Notes

[1] Operational scale scores have lower possible scale scores than those used in this study. Preliminary examinations of the cross-grade regressions indicated flat regressions for the very lowest scale scores, and therefore, the range of those lowest scale scores was constrained for all the analyses in this study.

[2] The regression techniques and multilevel models were applied to the scale scores that were the result of the vertical scaling process. However, these procedures can be applied without vertical scales. Given that the vertical scaling merely applies a linear transformation to each within-grade scale, the results of the regressions are, in essence, not affected by the existence of the vertical scale.

[3] It was assumed that nonmonotonicities that occurred were due to sampling effects, and that in potential operational uses of these regressions, it would not make sense to predict a lower 2006 score for students with higher 2005 scores. Therefore, the empirical regressions were smoothed to be nondecreasing.

[4] A random slope effect would have captured differences in the rate of change from 2005 to 2006 scores attributable to the school.

[5] A scaling factor can be used to constrain level-I dispersion. This was not done to maintain consistency with previous estimation methods used in this study.

[6] To explore the effects of retaining small schools in the analysis, models were re-estimated with small schools omitted. With the exclusion of small schools, the mean λ_j increased, as expected, but results were comparable in terms of fixed and random effects for remaining schools.