Measuring the Power of Learning.®

Research Report
ETS RR–17-54

Monitoring Score Change Patterns to Support *TOEIC*® Listening and Reading Test Quality

Youhua Wei

Albert Low

December 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Monitoring Score Change Patterns to Support *TOEIC*® Listening and Reading Test Quality

Youhua Wei[1] & Albert Low[2]

1 Educational Testing Service, Princeton, NJ
2 The Enrollment Management Association, Princeton, NJ

In most large-scale programs of tests that aid in making high-stakes decisions, such as the *TOEIC*® family of products and service, it is not unusual for a significant portion of test takers to retake the test at multiple times. The study reported here used multilevel growth modeling to explore the score change patterns of nearly 20,000 TOEIC Listening and Reading test takers who repeated the test six times during a 4-year period. The study revealed that (a) on average, repeaters' scores increased with each subsequent testing; (b) repeaters' score increases were larger for initial retests than for later ones; (c) test takers' educational backgrounds were related to their initial scores but not to their score increases; and (d) test takers' gender was related both to initial scores and to score increases. The results suggest that multilevel growth modeling analysis has potential for evaluating and monitoring test performance across administrations by exploring repeaters' score change patterns over time. The study also provided empirical evidence for the reliability and validity of TOEIC scores.

Keywords Repeater; growth modeling; quality control; TOEIC

doi:10.1002/ets2.12186

Quality control in educational measurement should be conducted systematically not only within individual administrations but also across administrations over time (von Davier, 2012). Across-administration test quality control may include the evaluation of the fluctuation of score summary statistics, population composition and background changes, test content evolution and difficulty shift, equating errors and scale drift, and the stability of psychometric properties such as reliability and validity. Various methods and procedures have been proposed for this purpose, such as time series analysis (Li, Li, & von Davier, 2011), harmonic regression (Lee & Haberman, 2013), linear mixed effects modeling (Lee, Liu, & von Davier, 2013), Shewhart control charts (see a brief description in von Davier, 2012), hidden Markov modeling (Lee & von Davier, 2013), and multilevel analysis (Wei, 2013; Wei & Qu, 2014).

In large-scale programs of tests that aid in making high-stakes decisions, some test takers take a test more than once, and they have been called *repeaters*. Repeater studies have been conducted to examine repeaters' score changes and explore their score growth patterns across administrations. Most studies (e.g., Kingston & Turner, 1984; Wei & Morgan, 2016; Yang, Bontya, & Moses, 2011; Zhang, 2008) have evaluated repeaters' score changes between two adjacent administrations; few studies (e.g., Nathan & Camara, 1998; Wilson, 1987) have explored repeaters' longitudinal score change patterns over multiple repetitions. On the basis of those studies, the average score changes tended to increase with the number of times tested, but the score changes were related to a number of factors, such as the number of repetitions, the interval between repetitions, initial scores, educational level, and gender.

The analyses based on the data collected at only two time points are often inadequate for investigating score growth. The longitudinal data tend to reduce quickly with the requirement of more retakes, so the results tend to be unstable. Typically, test takers may repeat a test a different number of times and at different points in time. Therefore repeaters' data tend to be unfixed and unbalanced, and advanced methods need to be used to take full advantage of the available information to provide a complete picture of repeaters' score growth trajectories, especially over a long time.

Multilevel growth modeling (e.g., Raudenbush & Bryk, 2002; Singer & Willett, 2003) is a flexible method that allows us to explore repeaters' longitudinal score change patterns when the number and spacing of time points vary across individual examinees. As for most testing programs, repeater data from the *TOEIC*® tests tend to be unbalanced and unfixed. That is, during any given period of time, test takers tend to retake the test different numbers of times and at variable intervals

Corresponding author: Y. Wei, E-mail: ywei@ets.org

between repetitions. Furthermore, time intervals tend to vary both within persons and between persons. Multilevel growth modeling can handle different data sets and fully use all repeaters' information to provide a more complete picture of repeaters' growth trajectories.

The repeaters' score changes over time can be used for quality control of test performance from different perspectives. First, because repeaters are the same examinees taking a test over time, their score changes can be used to evaluate the stability of test performance across administrations. A lack of stability can signal one or more of the issues mentioned at the beginning of the report. Second, repeaters' score changes provide empirical data to evaluate the score reliability by examining test score consistency across forms, across administrations, or over time based on the same examinees, especially when the intervals between repetitions are short. Third, repeaters' score changes provide operational data to evaluate score validity by comparing the growth patterns in a testing program with patterns found in other related testing programs or with related learning theories. Fourth, and finally, a testing program can use repeaters' growth patterns to predict and monitor their performance in future administrations.

The study reported here is based on repeaters' data from the TOEIC Listening and Reading test over a 4-year period from 2010 to 2014. Multilevel growth modeling was used to explore repeaters' test score change patterns. The growth modeling results were used for the quality control of test performance by evaluating the stability, reliability, and validity of test scores and the potential to monitor test performance across administrations.

Methodology

Data

The data were collected from the TOEIC Listening and Reading test in a country where English is a foreign language. The test has two sections, Listening and Reading, each consisting of 100 multiple-choice items. For each section, the raw scores range from 0 to 100 and the scale scores from 5 to 495 by increments of 5. Equating is conducted so that scale scores from different administrations or test forms are on the same scale. Therefore the longitudinal scale score data of the same test takers across administrations can be used to explore their score growth trajectories (Castellano & Ho, 2013). At each administration, a questionnaire is used to collect information on test takers' general background, English learning experience, and test-taking experience.

The test is offered in strictly scheduled monthly administrations in the country, with each administration using one unique test form. The data used in this study include Listening and Reading scale scores and background information of 19,855 test takers who had taken the test six times in 68 administrations in 4 years from 2010 to 2014. The spacing of test taking (in terms of months) varied across test takers. Table 1 shows the distribution of test takers based on the time gaps between adjacent times tested within the 4 years (e.g., between the first and second times and between the second and third times). The table shows that the number of repeaters tended to decrease when the time gap between adjacent repetitions became longer.

Data Preparation

As in a typical multilevel growth analysis (e.g., Raudenbush & Bryk, 2002; Singer & Willett, 2003), the repeated measures of each test taker in this study were considered as nested within the person. Therefore the repeaters' data had two levels, with repeated measures, including the scale scores and time-varying background, in multiple test-taking months as the Level 1 variables and unchanged person-level characteristics as the Level 2 variables.

At Level 1, the test taker's scale score in each of the multiple administrations was the dependent variable and the administration time was the independent variable. The Listening scale scores ranged from 105 to 495, with a mean of 334 and a standard deviation of 74. The Reading scale scores ranged from 85 to 495, with a mean of 279 and a standard deviation of 82. The administration time was defined as the amount of time in months that had elapsed from the first time a test taker took the test in the 4 years. The starting month and the spacing of the six test-taking months varied across test takers. For example, if one test taker took the test in January, March, May, August, November, and December in the first year, his or her administration times would be 0, 2, 4, 7, 10, and 11. If another test taker took the test in September and December in the first year, and then took the test in January, May, July, and October in the second year, his or her administration times would be 0, 3, 4, 8, 10, and 13. Therefore the possible administration times ranged from 0 to 47 in months in the 4 years.

Table 1 Distribution of Repeaters Based on Time Gaps Between Adjacent Times Tested

Time gap (month)	First – second	Second – third	Third – fourth	Fourth – fifth	Fifth – sixth
1	3,865	4,301	4,615	4,593	4,393
2	4,536	4,986	5,044	4,909	4,356
3	1,901	2,025	1,921	1,970	2,046
4	2,759	2,420	2,330	2,108	1,886
5	1,031	833	839	842	827
6	1,288	1,314	1,370	1,376	1,396
7	486	570	502	518	529
8	816	795	735	745	711
9	265	308	331	345	378
10	446	484	489	485	569
11	268	258	219	265	307
12	752	486	469	516	764
13	173	129	118	121	180
14	239	191	164	207	275
15	95	92	75	87	130
16	177	126	162	165	180
17	119	65	56	89	115
18	135	92	81	116	174
19	56	42	29	47	75
20	101	72	65	64	103
21	29	37	36	36	68
22	71	45	45	50	88
23	33	25	26	27	54
24	61	42	35	61	66
25	26	18	22	16	27
26	28	25	20	29	42
27	13	9	6	8	22
28	24	19	19	18	34
29	14	11	6	14	16
30	10	11	5	6	14
31	4	6	6	4	5
32	10	4	5	7	5
33	2	5	0	1	3
34	10	3	2	3	3
35	3	4	3	3	2
36	4	2	3	4	2
37	1	0	1	0	2
38	2	0	1	0	3
39	1	0	0	0	2
40	1	0	0	0	3

Note. $N = 19{,}855$.

Two types of test takers' background information tended to change across the six times of test taking and had close relations with test takers' scale scores. The first one was the test takers' *occupation status*, which was based on the survey question "Which of the following best describes your current status" (see Table 2 for the options); the second one was the test takers' *daily English use time*, which was based on the question "How much time must you use English in your daily life?" (see Table 2 for the options). These two background variables were selected and used as the time-varying independent variables or covariates at Level 1.

Test takers' *gender* information remained the same, and their *educational levels* tended to be the same or very similar across the six times of test taking (see Table 2 for the specific educational levels). The examinees' *test-taking experience* before the first time tested in the 4 years of data collection period was another type of background information. It was based on test takers' responses to the question "Before today, how many times have you taken the test?" at the first time tested in the 4 years (see Table 2 for the options). These three types of background information (i.e., gender, educational level, and test-taking experience) also had close relations with test takers' scale scores, and they were used as the unchanged person-level characteristics at Level 2.

Table 2 Variables and Codes at Levels 1 and 2

Data level	Variable	Options	Code	Subgroup percentage	Variable name
1	Current occupation	Full-time employed	(0, 0, 0, 0)	54.95	
		Missing information	(1, 0, 0, 0)	2.19	EMPMIS
		Part-time employed	(0, 1, 0, 0)	3.54	EMPPAR
		Unemployed	(0, 0, 1, 0)	3.77	UNEMP
		Full-time student	(0, 0, 0, 1)	35.56	STUDT
	Daily English use time	None at all	1	26.37	ENGUSE
		1% – 10% and missing information	2	47.54	
		11 – 20%	3	14.59	
		21 – 50%	4	9.37	
		51 – 100%	5	2.13	
	Time	$M = 11.55$, S.D. = 10.90, min. = 0, max. = 47			TIME
	Listening score	$M = 333.75$, S.D. = 74.23, min. = 105, max. = 495			LISTEN
	Reading score	$M = 279.30$, S.D. = 82.35, min. = 85, max. = 495			READ
2	Education level	Vocational/technical high school	(0, 0, 0, 0, 0, 0, 0, 0)	1.56	
		Missing information/primary school	(1, 0, 0, 0, 0, 0, 0, 0)	1.77	EDUMIS
		Junior high school	(0, 1, 0, 0, 0, 0, 0, 0)	0.11	SECOND1
		High school	(0, 0, 1, 0, 0, 0, 0, 0)	3.62	SECOND2
		Vocational/technical school	(0, 0, 0, 1, 0, 0, 0, 0)	2.08	VOTECH
		Community college	(0, 0, 0, 0, 1, 0, 0, 0)	3.00	COMMUN
		Undergraduate	(0, 0, 0, 0, 0, 1, 0, 0)	70.03	UNDERG
		Graduate	(0, 0, 0, 0, 0, 0, 1, 0)	17.63	GRADUA
		Language institute	(0, 0, 0, 0, 0, 0, 0, 1)	0.20	LANGUA
	Gender	Male	0	65.19	
		Female	1	34.81	GENDER
	Test-taking experience	Tested at least once before	0	77.65	
		Never tested before	1	22.35	PREEXP

Among the five background variables, occupation status, gender, educational level, and test-taking experience were categorical, so dummy coding was conducted for these four background variables. The daily English use time was ordinal, so a Likert scale was used to quantify its values. Table 2 shows the background variables and their codes at Levels 1 and 2. The coding was mainly based on the survey questions' original response options. The test performance patterns of subgroups based on response options were also taken into account for the coding. For example, based on the survey question about test takers' occupation status, the subgroup with missing information tended to have consistent performance compared with other subgroups, so this subgroup was not removed from the sample but rather was coded as a separate subgroup. On the basis of the survey question about test takers' daily English use time, the subgroup who chose "1 – 10%" and the subgroup who did not choose any option tended to have similar test performance, so these two subgroups were combined and coded as one subgroup for analysis. For the convenience of interpretation, the subgroups with lowest test performance were coded as the reference groups in most cases. For example, the subgroup of full-time employed for the occupation status background was coded as the reference group; the subgroup choosing the option of vocational/technical high school for the educational level was coded as the reference group.

Preliminary Analyses

Some descriptive analyses were conducted to explore the nature and idiosyncrasies of the repeaters' growth trajectories before the multilevel growth modeling was used. On the basis of the observation of some randomly selected repeaters' scores across repetitions, the scores tended to increase over time, but the rate of increase slowed gradually, with a substantial variation across individuals. Although the starting month and the spacing of the six test-taking months varied across test takers, each repeater had scores at six time points in the 0 – 47 administration months over 4 years. To show the score change trend at the group level over time, we computed repeaters' scale score means at each of the 48 time points based on the data available at each administration time, and then plotted the score means over time (i.e., months). Figure 1 shows the plots of the observed score means for Listening and Reading. The plots show that repeaters' scale score means tended to increase over time, but the increasing rate tended to decrease gradually. Therefore the preliminary analyses based on both

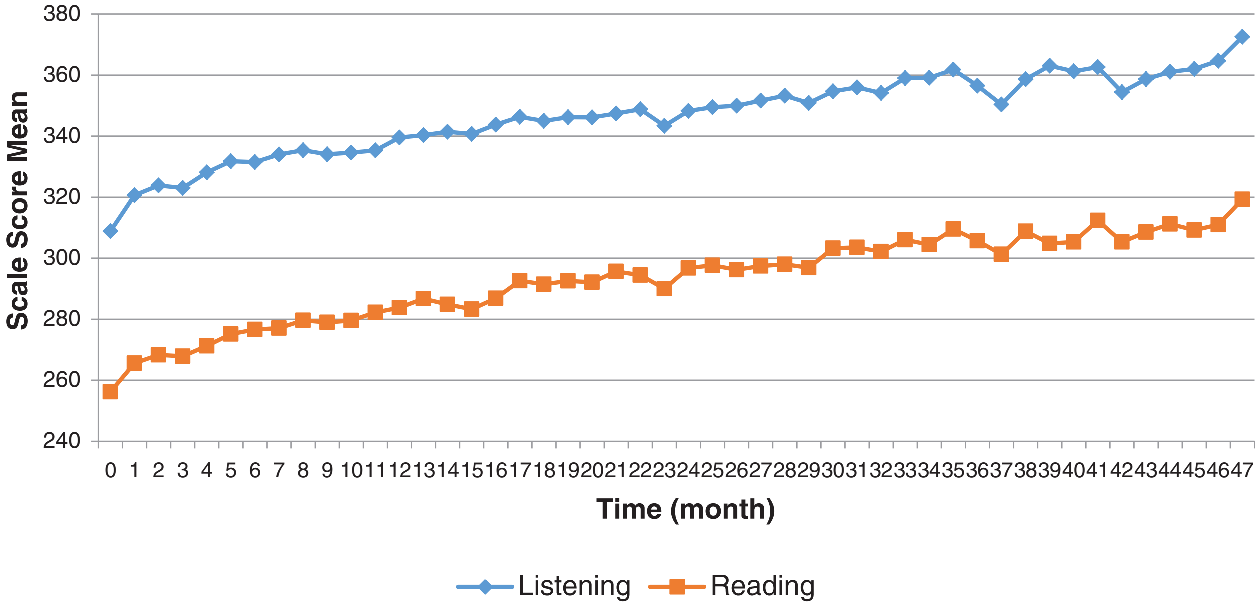

Figure 1 Listening and Reading observed score means over time (month).

individual and group data suggest a nonlinear growth model for repeaters' score changes for both Listening and Reading. The relationships between test takers' scale scores and their background variables were also explored in the preliminary analyses.

Multilevel Growth Modeling

Multilevel growth modeling was used to explore the repeaters' score change patterns, with examinees' repeated measures at Level 1 and person-level characteristics at Level 2. On the basis of the preliminary analyses of repeaters' score changes and the relations between examinees' scores and their background information, different models were explored and results were evaluated in terms of model fit, growth parameter estimation, variance estimation, and test performance prediction.

Following the suggestions by Raudenbush and Bryk (2002) and Singer and Willett (2003) on model building, the analyses started with simple growth models and then used the step-up strategy to include more growth parameters and background variables based on promising submodels. Specifically, four types of models were used in this study (see the appendix for the statistical specifications of the four models).

Unconditional Means Model

As the simplest model, the unconditional means model does not include any predictors and does not describe the score change over time. However, this model partitions the total score variation into the within-person variation at Level 1 and the between-person variation at Level 2. It helps determine whether there is sufficient variation to warrant further analysis at each level.

Linear Growth Model

This model includes the linear TIME predictor in the Level 1 model. Assuming a constant rate of score change over time, this model estimates the repeaters' average score change per month. It also estimates the between-person variation in the rate of score change.

Quadratic Growth Model

Assuming the rate of score change is not constant over time, this model includes both the linear TIME predictor and the quadratic TIME^2 predictor in the Level 1 model. The linear growth parameter estimates the instantaneous or initial rate of change. The quadratic parameter estimates the acceleration in the growth trajectory.

Conditional Quadratic Growth Model

This model includes test takers' background variables in the quadratic growth model, so that the impact of examinees' background on their score growth trajectory can be examined.

The analyses first explored the repeaters' growth trajectories by evaluating different growth models without including any background variables. When necessary, polynomial models with higher degrees (e.g., cubic growth model by including the cubic TIME^3 predictor) were explored and examined. After the most appropriate growth model was identified, the time-varying background variables were added in the Level 1 model, and the person-level background variables were added in the Level 2 models, so their impacts on examinees' test scores and growth parameters could be examined.

Model Validation

The data of the other 1,861 examinees who had taken the test 12 times in the same 4 years were used to validate the repeaters' growth models, which were selected based on the original data of 19,855 examinees. The models, parameter estimates, and the impacts of background variables were compared to evaluate the validity of the models.

Results

In this section, we first summarize the results from the unconditional means model, which can provide baseline information for further analysis. Then we present the modeling results based on the linear, quadratic, and cubic growth models, followed by the results from the conditional growth model, which include background variables in both Level 1 and Level 2 models. We close the section by evaluating the validity of the model identified from the study.

Unconditional Means Model

Listening

On the basis of the unconditional means model for Listening scores (see Table 3), the estimated grand mean of all repeaters' scores across the six administration times in the study was 333.75. The Level 1 variance estimate was 1,242.78, and the Level 2 variance estimate was 4,267.24, which indicates that much more score variation came from the between-person variation (77%) than the over-time within-person variation (23%). However, both the within-person variation (S.D. = 35.25) and the between-person variation (S.D. = 65.32) in the test scores were large enough to warrant further analysis. Therefore predictors at both levels would be necessary to explore the variation of the test scores.

Reading

On the basis of the unconditional means model for Reading scores (see Table 4), the grand mean estimate was 279.30, and 81% of the test score variation came from the between-person variation. Both the within-person variation (S.D. = 36.17) and the between-person variation (S.D. = 73.98) in Reading scores were large enough to warrant further analysis.

These results were based on the unconditional means model with the assumption of homogeneity of Level 1 variance across times. The likelihood ratio test suggests that the Level 1 variance was not homogeneous for both Listening and Reading scores. However, the estimation of fixed effects and their standard errors was robust to the violation of this assumption (Kasim & Raudenbush, 1998). A general estimate of Level 1 variance was needed in further analysis to estimate the variance explained by Level 1 predictors. Therefore results from the unconditional means model with the assumption of homogeneity of Level 1 variance were used for this study. Although both the within-person variation and the between-person variation in the test scores were sufficient to warrant further analysis, we first focused on the within-person variation by including growth parameters and time-varying background variables in the Level 1 model in the following analyses.

Table 3 Results From Unconditional Means Model for Listening: $Y_{ti} = \pi_{0i} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}$

Fixed	Coefficient	S.E.	*t*-Ratio	*df*	*p*
Grand mean	333.75	0.47	703.08	19,854	0.00
Random	Variance component	S.D.	Chi-square	*df*	*p*
Person-specific mean	4,267.24	65.32	428,882.10	19,854	0.00
Level 1 error	1,242.78	35.25			
Model fit	Deviance		Parameters		
	1,247,896.79		2		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	40,829.70		19,854	0.00	

Table 4 Results From Unconditional Means Model for Reading: $Y_{ti} = \pi_{0i} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}$

Fixed	Coefficient	S.E.	*t*-Ratio	*df*	*p*
Grand mean	279.30	0.54	521.66	19,854	0.00
Random	Variance component	S.D.	Chi-square	*df*	*p*
Person-specific mean	5,473.74	73.98	518,221.43	19,854	0.00
Level 1 error	1,308.38	36.17			
Model fit	Deviance		Parameters		
	1,257,781.68		2		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	36,086.43		19,854	0.00	

Linear Growth Model

Listening

On the basis of the results from the linear growth model for Listening scores (see Table 5), the repeaters' average initial score at the first administration time was 316.50, and their scores increased on average by 1.58 points per month in the 4 years of the data collection period. However, there were substantial between-individual variations in both the initial status and increase rate. Specifically, 95% of the repeaters' initial scores were in the range of $316.50 \pm 1.96 * \sqrt{4592.59} = (183.67, 449.33)$, and 95% of the repeaters' score growth rates were in the range of $1.58 \pm 1.96 * \sqrt{1.49} = (-.81, 3.97)$. In addition, there was a slight negative correlation (−.26) between examinees' initial status and growth rate. Compared with the unconditional means model, the estimated Level 1 residual variance decreased by $(1{,}242.78 - 920.26)/1{,}242.78 = 26\%$. Therefore 26% of the within-person variation in Listening scores was associated with the linear TIME, but a substantial amount of variance still remained unexplained at Level 1, and more predictors needed to be included in the Level 1 model.

Reading

On the basis of the linear growth model for Reading scores (see Table 6), the repeaters' average initial score was 262.33, with 95% of the initial scores in the range of (115.76, 408.90); the repeaters' scores increased on average by 1.55 points per month, with 95% of the growth rates in the range of (−.88, 3.98). A slight negative correlation (−.15) between examinees' initial score and growth rate was also found in Reading scores. Compared with the unconditional means model, about 25% of the within-person variation in Reading scores was associated with the linear TIME. Again, a substantial

Table 5 Results From Linear Growth Model for Listening: $Y_{ti} = \pi_{0i} + \pi_{1i}\text{TIME}_{ti} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}$, $\pi_{1i} = \beta_{10} + r_{1i}$

Fixed	Coefficient	S.E.	*t*-Ratio	*df*	*p*
Mean initial status	316.50	0.50	629.93	19,854	0.00
Mean growth rate	1.58	0.01	112.13	19,854	0.00
Random	Variance component	S.D.	Chi-square	*df*	*p*
Initial status	4,592.59	67.77	227,324.50	19,854	0.00
Growth rate	1.49	1.22	39,265.38	19,854	0.00
Level 1 error	920.26	30.34			
Model fit	Deviance		Parameters		
	1,228,466.99		4		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	33,089.66		19,844	0.00	
Tau as correlations	Initial status		Growth rate		
Initial status	1				
Growth rate	−0.26		1		

Table 6 Results From Linear Growth Model for Reading: $Y_{ti} = \pi_{0i} + \pi_{1i}\text{TIME}_{ti} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}$, $\pi_{1i} = \beta_{10} + r_{1i}$

Fixed	Coefficient	S.E.	*t*-Ratio	*df*	*p*
Mean initial status	262.33	0.55	475.53	19,854	0.00
Mean growth rate	1.55	0.01	106.47	19,854	0.00
Random	Variance component	S.D.	Chi-square	*df*	*p*
Initial status	5,591.54	74.78	252,925.56	19,854	0.00
Growth rate	1.54	1.24	38,419.80	19,854	0.00
Level 1 error	983.37	31.36			
Model fit	Deviance		Parameters		
	1,239,485.53		4		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	31,766.17		19,854	0.00	
Tau as correlations	Initial status		Growth rate		
Initial status	1				
Growth rate	−0.15		1		

amount of variance still remained unexplained at Level 1, and more predictors needed to be included in the Level 1 model.

Quadratic Growth Model

Listening

On the basis of the quadratic growth model for Listening scores (see Table 7), on average, the estimated initial score was 312.16, the initial growth rate was 2.83, and acceleration was −.04. The statistically significant negative mean acceleration indicates that repeaters improved their scores at a decreasing rate over time. However, there still were substantial interindividual variations in the initial score, initial growth rate, and acceleration, with 95% of the growth parameters in the ranges of (178.90, 445.42), (−1.89, 7.55), and (−.13, −.05), respectively. There was a slight negative correlation (−.20) between examinees' initial status and initial growth rate but a strong negative correlation between initial growth rate and

Table 7 Results From Quadratic Growth Model for Listening: $Y_{ti} = \pi_{0i} + \pi_{1i}\text{TIME}_{ti} + \pi_{2i}\text{TIME}^2_{ti} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}$, $\pi_{1i} = \beta_{10} + r_{1i}$, $\pi_{2i} = \beta_{20} + r_{2i}$

Fixed	Coefficient	S.E.	*t*-Ratio	*df*	*p*
Mean initial status	262.33	0.55	475.53	19,854	0.00
Mean growth rate	1.55	0.01	106.47	19,854	0.00
Mean acceleration	−0.04	0.00	−49.43	19,854	0.00
Random	Variance component	S.D.	Chi-square[a]	*df*[a]	*p*[a]
Initial status	4,622.63	67.99	110,895.78	14,007	0.00
Initial growth rate	5.81	2.41	16,482.67	14,007	0.00
Acceleration	0.00	0.05	15,732.26	14,007	0.00
Level 1 error	874.44	29.57			
Model fit	Deviance		Parameters		
	1,225,381.39		7		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	22,600.47		13,998	0.00	
Tau as correlations	Initial status	Initial growth rate	Acceleration		
Initial status	1				
Initial growth rate	−0.20	1			
Acceleration	0.10	−0.95	1		

[a] The chi-square statistics are based on 14,008 of 19,855 units that had sufficient data for computation.

Table 8 Results From Quadratic Growth Model for Reading: $Y_{ti} = \pi_{0i} + \pi_{1i}\text{TIME}_{ti} + \pi_{2i}\text{TIME}^2_{ti} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}$, $\pi_{1i} = \beta_{10} + r_{1i}$, $\pi_{2i} = \beta_{20} + r_{2i}$

Fixed	Coefficient	S.E.	*t*-Ratio	*df*	*p*
Mean initial status	258.74	0.56	463.17	19,854	0.00
Mean growth rate	2.59	0.03	78.74	19,854	0.00
Mean acceleration	−0.04	0.00	−39.46	19,854	0.00
Random	Variance component	S.D.	Chi-square[a]	*df*[a]	*p*[a]
Initial status	5,605.55	74.87	123,840.97	14,007	0.00
Initial growth rate	6.17	2.48	16,436.96	14,007	0.00
Acceleration	0.00	0.05	15,575.27	14,007	0.00
Level 1 error	939.35	30.65			
Model fit	Deviance		Parameters		
	1,237,339.16		7		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	22,349.37		14,007	0.00	
Tau as correlations	Initial status	Initial growth rate	Acceleration		
Initial status	1				
Initial growth rate	−0.12	1			
Acceleration	0.05	−0.94	1		

[a] The chi-square statistics are based on 14,008 of 19,855 units that had sufficient data for computation.

acceleration (−.95). Compared with the linear growth model, 4% more within-person variation in Listening scores was associated with the addition of the quadratic parameter. However, a substantial amount of variance was unpredicted at Level 1.

Table 9 Results From Cubic Growth Model for Listening: $Y_{ti} = \pi_{0i} + \pi_{1i}\text{TIME}_{ti} + \pi_{2i}\text{TIME}_{ti}^{2} + \pi_{3i}\text{TIME}_{ti}^{3} + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}$, $\pi_{1i} = \beta_{10} + r_{1i}$, $\pi_{2i} = \beta_{20} + r_{2i}$, $\pi_{3i} = \beta_{30} + r_{3i}$

Fixed	Coefficient	S.E.	*t*-Ratio	*df*	*p*
Mean initial status	310.44	0.52	602.45	19,854	0.00
Mean linear	3.91	0.06	67.14	19,854	0.00
Mean quadratic	−0.13	0.00	−33.19	19,854	0.00
Mean cubic	0.00	0.00	23.32	19,854	0.00
Random	Variance component	S.D.	Chi-square[a]	*df*[a]	*p*[a]
Initial status	4,632.30	68.06	9,777.31	1,671	0.00
Linear	12.25	3.50	1,597.36	1,671	>0.50
Quadratic	0.03	0.18	1,572.82	1,671	>0.50
Cubic	0.00	0.00	1,583.87	1,671	>0.50
Level 1 error	853.96	29.22			
Model fit	Deviance		Parameters		
	1,224,738.20		11		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	2,841.46		1,671	0.000	
Tau as correlations	Initial status	Linear	Quadratic	Cubic	
Initial status	1				
Linear	−0.17	1			
Quadratic	0.07	−0.90	1		
Cubic	−0.05	0.79	−0.98	1	

[a] The chi-square statistics are based on 1,672 of 19,855 units that had sufficient data for computation.

Reading

On the basis of the quadratic growth modeling results for Reading scores (Table 8), the estimated initial score, initial growth rate, and acceleration were 258.74, 2.59, and −.04, with 95% of the growth parameters in the ranges of (111.99, 405.49), (−2.28, 7.46), and (−.14, .06), respectively. A slight negative correlation between examinees' initial status and initial growth rate (−.12) and a strong negative correlation between initial growth rate and acceleration (−.94) were also found in Reading scores. Compared with the linear growth model, 3% more within-person variation in Reading scores was associated with the addition of the quadratic parameter. Again, there was still a substantial amount of unexplained variance at Level 1.

Cubic Growth Model

To explore the repeaters' score growth trajectories, the cubic growth parameter TIME^{3} was added in the Level 1 quadratic growth model. The results from the cubic growth models for Listening and Reading scores are presented in Tables 9 and 10, respectively. Compared with the quadratic growth modeling results, 1% and 2% more within-person variation in Listening and Reading scores, respectively, was associated with the addition of the cubic parameter.

On the basis of the cubic model, all the growth parameters, except for the initial score, were fixed with no interindividual variations for both Listening and Reading scores, which is not consistent with what we found in the preliminary analyses and other models. It seems implausible that the linear, quadratic, and cubic growth parameters were invariant across individual examinees. It is very possible that there were not enough data for the cubic model to produce accurate chi-square statistics for the variances of the parameter estimates (e.g., the chi-square statistics were based on only 1,672 of 19,855 examinees). On the basis of the model fit statistics, the deviance drop was substantially smaller than the deviance drop from the linear growth model to the quadratic model and the deviance drop from the linear growth model to the unconditional means model.

Table 10 Results From Cubic Growth Model for Reading: $Y_{ti} = \pi_{0i} + \pi_{1i}\text{TIME}_{ti} + \pi_{2i}\text{TIME}_{ti}^2 + \pi_{3i}\text{TIME}_{ti}^3 + e_{ti}$, $\pi_{0i} = \beta_{00} + r_{0i}$, $\pi_{1i} = \beta_{10} + r_{1i}$, $\pi_{2i} = \beta_{20} + r_{2i}$, $\pi_{3i} = \beta_{30} + r_{3i}$

Fixed	Coefficient	S.E.	*t*-Ratio	*df*	*p*
Mean initial status	257.28	0.56	456.44	19,854	0.00
Mean linear	3.51	0.06	57.70	19,854	0.00
Mean quadratic	−0.11	0.00	−27.45	19,854	0.00
Mean cubic	0.00	0.00	19.84	19,854	0.00
Random	Variance component	S.D.	Chi-square[a]	*df*[a]	*p*[a]
Initial status	5,620.28	74.97	11,601.14	1,671	0.00
Linear	14.09	3.75	1,640.27	1,671	>0.50
Quadratic	0.03	0.18	1,597.81	1,671	>0.50
Cubic	0.00	0.00	1,585.51	1,671	>0.50
Level 1 error	918.68	30.31			
Model fit	Deviance		Parameters		
	1,236,790.95		11		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	2,895.64		1,671	0.00	
Tau as correlations	Initial status	Linear	Quadratic	Cubic	
Initial status	1				
Linear	−0.12	1			
Quadratic	0.06	−0.92	1		
Cubic	−0.05	0.86	−0.99	1	

[a]The chi-square statistics are based on 1,672 of 19,855 units that had sufficient data for computation.

To further evaluate model fit, at each of the 0–47 administration time points, we computed repeaters' fitted score means based on individual growth trajectories and then plotted the fitted score means over time based on both quadratic and cubic models. Figures 2 and 3 show the plots of observed score means and fitted score means based on the two models for Listening and Reading, respectively. On the basis of the plots, the two models had very similar model fit with the observed data, except at the last four time points, where the cubic growth trajectories were slightly better than the quadratic growth trajectories.

On the basis of the growth parameter estimates, fit statistics, and the principle of parsimony in statistical modeling, the quadratic growth model was considered to be the appropriate growth model in this study.

Conditional Quadratic Growth Model

The quadratic growth model was selected to describe repeaters' score change patterns over time for both Listening and Reading. However, the examinees' background might have an impact on their score change patterns. Therefore two time-varying background variables, current occupation and daily English use time, were added in the Level 1 quadratic growth model, and three person-level background variables, gender, educational level, and test-taking experience, were added as predictors in the Level 2 models for the growth parameters. The two time-varying background variables were first separately added in the Level 1 model and their impacts on examinees' scores over time were evaluated; after the important time-varying background variables were selected to remain in the Level 1 model, the three person-level background variables were added in the Level 2 models to evaluate their impacts on the growth parameters.

Listening

On the basis of the model fit statistics, fixed effect coefficients, and the principle of parsimony in statistical modeling, the time-varying background variable daily English use time was selected to remain in the Level 1 quadratic model; the person-level background variables gender and test-taking experience remained in the Level 2 models for all three growth

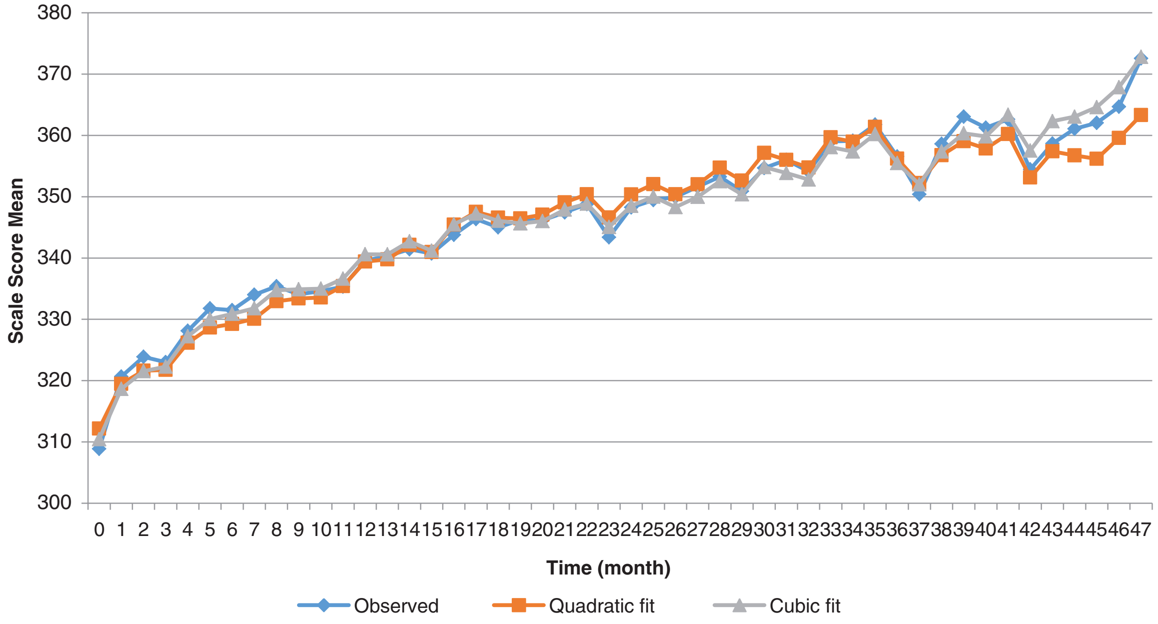

Figure 2 Fitted and observed score means over time for Listening.

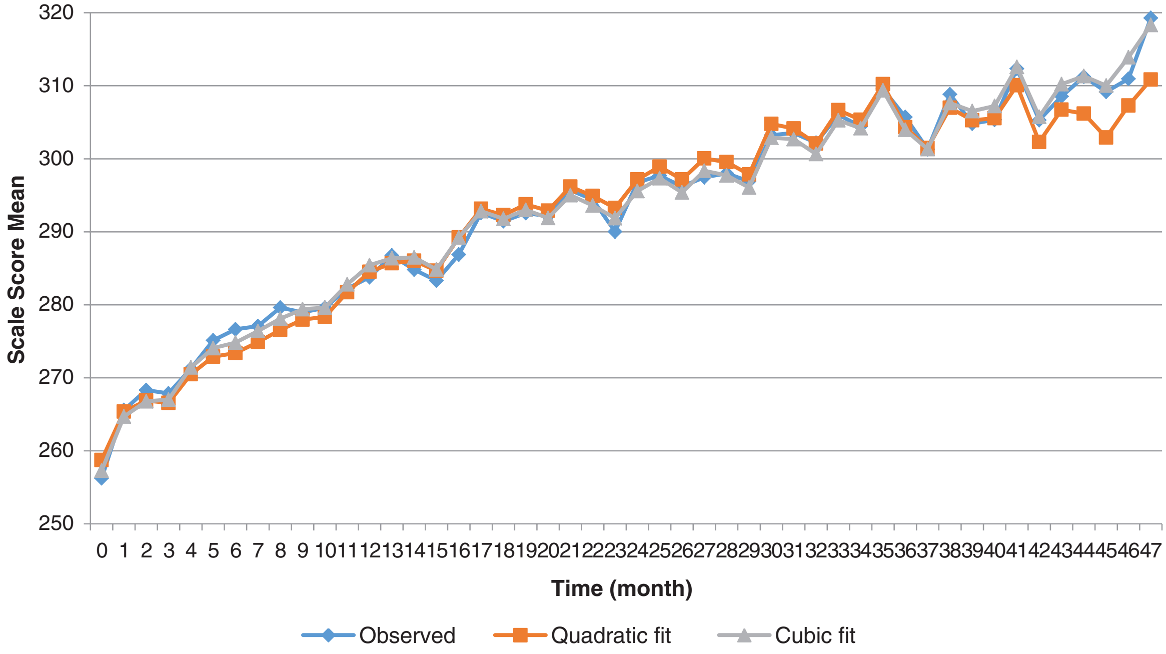

Figure 3 Fitted and observed score means over time for Reading.

parameters (i.e., initial status, initial growth rate, and acceleration), and educational level remained in the Level 2 model only for the initial status.

Table 11 shows the results from the final conditional quadratic model for Listening scores. On the basis of the model, the examinees' background variables had significant impacts on their Listening score growth trajectories. For example, for the initial status, women had higher average scores than men; examinees with vocational/technical high school education had lower average score than examinees with all other educational levels; and examinees without previous test-taking experience had lower average scores than examinees with experience. The examinees' scores tended to increase with their daily English use time. For score growth rate, women had lower initial growth rates than men; examinees without

Table 11 Results From Conditional Quadratic Model for Listening: $Y_{ti} = \pi_{0i} + \pi_{1i}\text{ENGUSE}_{ti} + \pi_{2i}\text{TIME}_{ti} + \pi_{3i}\text{TIME}^2_{ti} + e_{ti}$, $\pi_{0i} = \beta_{00} + \beta_{01}\text{GENDER}_i + \beta_{02}\text{EDUMIS}_i + \beta_{03}\text{SECOND1}_i + \beta_{04}\text{SECOND2}_i + \beta_{05}\text{VOTECH}_i + \beta_{06}\text{COMMUN}_i + \beta_{07}\text{UNDERGR}_i + \beta_{08}\text{GRADUA}_i + \beta_{09}\text{LANGUA}_i + \beta_{10}\text{PREEXP}_i + r_{0i}$, $\pi_{1i} = \beta_{10}$, $\pi_{2i} = \beta_{20} + \beta_{21}\text{GENDER}_i + \beta_{22}\ \text{PREEXP}_i + r_{2i}$, $\pi_{3i} = \beta_{30} + \beta_{31}\text{GENDER}_i + \beta_{32}\text{PREEXP}_i + r_{3i}$

Fixed	Coefficient	S.E.	*t*-Ratio	*df*	*p*
For initial status					
Intercept	271.91	3.40	79.91	19,844	0.00
GENDER	30.87	1.06	29.08	19,844	0.00
EDUMIS	35.68	4.87	7.33	19,844	0.00
SECOND1	74.87	13.49	5.55	19,844	0.00
SECOND2	22.94	4.39	5.23	19,844	0.00
VOTECH	11.86	4.81	2.47	19,844	0.01
COMMUN	33.13	4.51	7.35	19,844	0.00
UNDERGR	32.58	3.43	9.51	19,844	0.00
GRADUA	32.87	3.54	9.29	19,844	0.00
LANGUA	57.73	10.05	5.75	19,844	0.00
PREEXP	−34.18	1.13	−30.14	19,844	0.00
For ENGUSE slope					
Intercept	2.74	0.18	15.60	119,112	0.00
For mean linear slope					
Intercept	2.63	0.04	63.08	19,852	0.00
GENDER	−0.42	0.06	−6.47	19,852	0.00
PREEXP	1.46	0.08	17.22	19,852	0.00
For mean acceleration slope					
Intercept	−0.04	0.00	−35.43	19,852	0.00
GENDER	0.01	0.00	4.68	19,852	0.00
PREEXP	−0.02	0.00	−9.10	19,852	0.00
Random	Variance component	S.D.	Chi-square[a]	*df*[a]	*p*[a]
Initial status	4,104.09	64.06	100,977.36	13,997	0.00
Initial growth rate	5.51	2.35	16,385.22	14,005	0.00
Acceleration	0.00	0.05	15,741.04	14,005	0.00
Level 1 error	874.24	29.57			
Model fit	Deviance		Parameters		
	1,222,947.36		7		
Test of homogeneity of Level 1 variance	Chi-square		*df*	*p*	
	22,886.25		14,005	0.00	
Tau as correlations	Initial status	Initial growth rate		Acceleration	
Initial status	1				
Initial growth rate	−0.14	1			
Acceleration	0.03	−0.95		1	

[a] The chi-square statistics are based on 14,008 of 19,855 units that had sufficient data for computation.

test-taking experience had higher initial growth rate than examinees with experience; and both gender and test-taking experience had impacts on the acceleration of the score growth.

Reading

Similar to the modeling results for Listening scores, the variable daily English use time was selected to remain in the Level 1 quadratic model for Reading scores. However, the variable gender remained only in the models for the initial status and initial growth rate, the variable educational level remained only in the model for the initial status, and the variable test-taking experience remained in the models for all three growth parameters. Table 12 shows the results from the final conditional quadratic model for Reading scores. Compared with the findings for Listening scores, the examinees'

Table 12 Results From Conditional Quadratic Model for Reading: $Y_{ti} = \pi_{0i} + \pi_{1i}\text{ENGUSE}_{ti} + \pi_{2i}\text{TIME}_{ti} + \pi_{3i}\text{TIME}^2_{ti} + e_{ti}$, $\pi_{0i} = \beta_{00} + \beta_{01}\text{GENDER}_i + \beta_{02}\text{EDUMIS}_i + \beta_{03}\text{SECOND2}_i + \beta_{04}\text{COMMUN}_i + \beta_{05}\text{UNDERGR}_i + \beta_{06}\text{GRADUA}_i + \beta_{07}\text{LANGUA}_i + \beta_{08}\text{PREEXP}_i + r_{0i}$, $\pi_{1i} = \beta_{10}$, $\pi_{2i} = \beta_{20} + \beta_{21}\text{GENDER}_i + \beta_{22}\text{PREEXP}_i + r_{2i}$, $\pi_{3i} = \beta_{30} + \beta_{31}\text{PREEXP}_i + r_{3i}$

Fixed	Coefficient	S.E.	*t*-Ratio	*df*	*p*
For initial status					
Intercept	202.73	2.72	74.61	19,846	0.00
GENDER	6.48	1.15	5.63	19,846	0.00
EDUMIS	56.72	4.78	11.88	19,846	0.00
SECOND2	11.44	3.95	2.89	19,846	0.00
COMMUN	39.57	4.25	9.32	19,846	0.00
UNDERGR	57.19	2.71	21.08	19,846	0.00
GRADUA	65.52	2.93	22.37	19,846	0.00
LANGUA	47.84	11.68	4.10	19,846	0.00
PREEXP	−27.46	1.27	−21.65	19,846	
For ENGUSE slope					
Intercept	2.72	0.18	14.97	119,115	0.00
For mean linear slope					
Intercept	2.31	0.04	62.49	19,852	0.00
GENDER	−0.15	0.03	−5.34	19,852	0.00
PREEXP	1.36	0.09	15.71	19,852	0.00
For mean acceleration slope					
Intercept	−0.03	0.00	−31.15	19,853	0.00
PREEXP	−0.02	0.00	−8.70	19,853	0.00
Random	**Variance component**	**S.D.**	**Chi-square[a]**	***df*[a]**	***p*[a]**
Initial status	5,176.85	71.95	115,746.18	13,999	0.00
Initial growth rate	5.73	2.39	16,283.00	14,005	0.00
Acceleration	0.00	0.05	15,518.41	14,006	0.00
Level 1 error	940.11	30.66			
Model fit	**Deviance**		**Parameters**		
	1,235,544.37		7		
Test of homogeneity of Level 1 variance	**Chi-square**		***df***	***p***	
	22,346.17		14,007	0.00	
Tau as correlations	**Initial status**	**Initial growth rate**		**Acceleration**	
Initial status	1				
Initial growth rate	−0.08	1			
Acceleration	0.02	−0.94		1	

[a] The chi-square statistics are based on 14,008 of 19,855 units that had sufficient data for computation.

background variables had similar impacts on their Reading score growth trajectories, except that gender did not have an impact on the acceleration of the Reading score growth.

For both Listening and Reading scores, the conditional quadratic growth model fit better than the quadratic growth model based on the deviance statistics. However, the examinees' initial growth rates still had very strong negative relations with acceleration over time (−.95 for Listening and −.94 for Reading). In addition, there were still substantial between-individual variations in the growth trajectories, which include examinees' initial status, initial growth rate, and acceleration over time.

Model Validation

The data of another group of 1,861 examinees who had taken the test 12 times in the same 4 years were used to examine the validity of the selected quadratic growth models (results are not presented in this report). A comparison of the linear, quadratic, and cubic growth models for the new group's Listening and Reading scores found that the quadratic model was the most appropriate model. The average initial score, initial growth rate, and acceleration based on 1,861 examinees were consistent with the growth parameters based on the 19,855 examinees. The strong negative correlation between initial growth rate and acceleration also remained consistent between the two samples.

When the conditional quadratic growth models based on the data of the 19,855 examinees were applied to the new group's Listening and Reading scores, the association of examinees' background variables with their growth trajectories remained similar.

Discussion

It is an important part of quality control for a testing program to monitor test performance across administrations, and various methods and procedures have been proposed for this purpose (von Davier, 2012). The existence of the same examinees who repeat the test in different administrations provides data to evaluate test performance over time. A testing program can use repeaters' data across administrations to examine score change patterns and then use these patterns to monitor test performance over time. This study used multilevel growth modeling to analyze a balanced but unfixed data set in which all examinees repeated the same number of test administrations but with variable intervals between test takings. The definition of TIME as the number of months that had elapsed from the first time tested and the use of equated scores from different administrations and forms put all examinees in the same framework for growth modeling analysis.

On the basis of the unconditional means model, the test scores varied much more among different examinees than they varied over time within persons. The within-person score variation in the 4 years was close to the standard error of score difference (i.e., 35; see Educational Testing Service [ETS], 2013) for each of the Listening and Reading sections in the TOEIC Listening and Reading test, which indicates the stability of test performance over time.

On the basis of the linear growth modeling results, the constant growth rate over time for both Listening and Reading was small, with about a 1.6 score point increase per month, which suggests the stability of repeaters' scores over time. However, as expected, the between-person variations in both initial status and growth rate were large: Individuals began at different proficiency levels, and they made progress at different rates. Although the linear growth model fit much better than the unconditional means model, a substantial proportion of within-person score variation still remained unrelated to the linear TIME predictor. In addition, a closer look at the plots of observed score means and fitted score means suggested that repeaters' scores did not increase at a constant rate, especially at the very beginning (i.e., from the first to the second times) and later times of testing. Therefore the linear growth model might describe repeaters' growth trajectories in the earlier times of testing (except for the first repetition) but may not account for the changing growth rate in their long-term score change patterns.

The quadratic growth model uses a linear parameter to estimate the initial growth rate at the very beginning and a quadratic parameter to estimate acceleration over time. The quadratic modeling results indicate that the repeaters' scores tended to increase more in their earlier repetitions, but the increase rate declined gradually over time. The repeater's growth trajectories based on the data in this study were consistent with the repeater score change patterns found in other testing programs, such as the *TOEFL*® test (Wilson, 1987), the *SAT*® I test (Nathan & Camara, 1998), and the *GRE*® General test (Rock & Werts, 1979).

The quadratic growth modeling yielded slight negative correlations between examinees' initial status and initial growth rate, which means that examinees with lower initial scores tended to have somewhat higher growth rates in their early times of testing. This finding was consistent with previous studies (Nathan & Camara, 1998; Wei & Morgan, 2016; Wilson, 1987; Yang et al., 2011). The relatively low negative correlations may be related to the repeater group composition in this study. Only 22.35% of the repeaters had never taken the test before. In other words, the majority of repeaters had taken the test at least once before the data collection period for this study, and the initial scores were not really their first-time scores in their test-taking experience. Accordingly, their initial scores in the data collection period did not show strong relations with their score changes. However, the quadratic growth modeling produced a very strong negative correlation between repeaters' initial growth rate and acceleration, which means that the examinees with lower initial growth rate tended to have higher acceleration over the growth trajectory. The quadratic growth modeling can easily find this repeater score change pattern, but descriptive and simple analyses often ignored the pattern.

The negative acceleration parameter estimate in the quadratic growth model suggests that the increase rate would decline over time. At what point in time would scores no longer exhibit any significant increase? The repeaters' fitted mean score plots based on individual growth trajectories did not show how the score increase rate changed over time. Figures 4 and 5 show the average quadratic and cubic growth functions (i.e., group growth trajectories) and observed score means

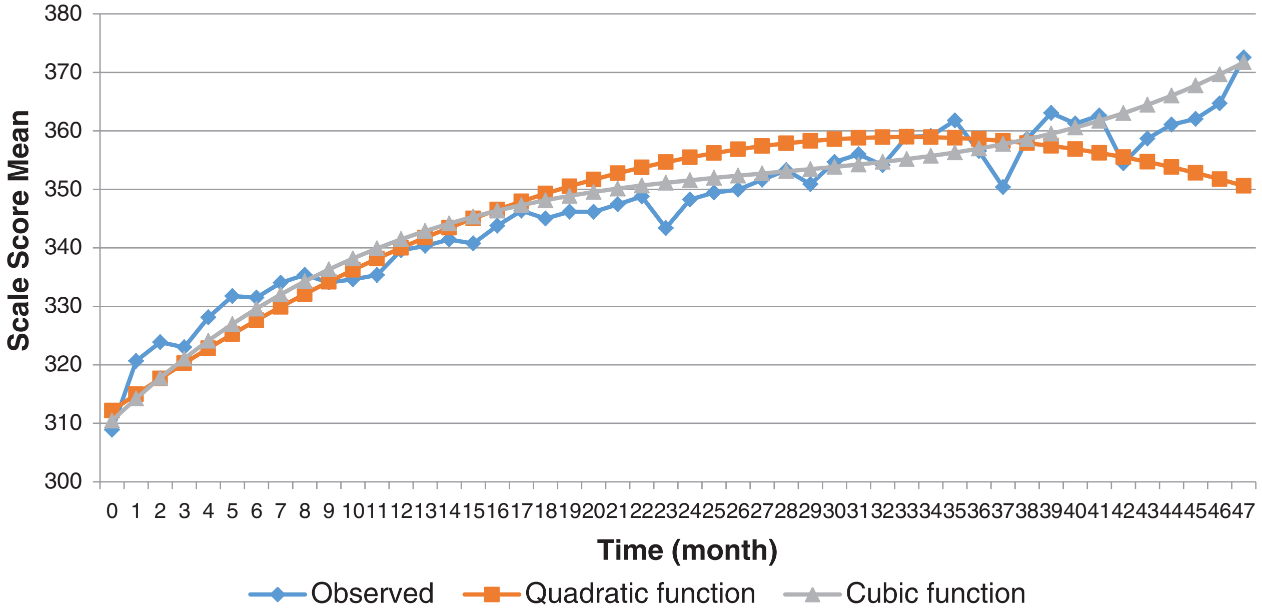

Figure 4 Growth functions and observed score means over time for Listening.

over time for Listening and Reading, respectively. The plots based on the quadratic growth function showed that the mean score growth rate changed from positive to negative in the 34th month for Listening and in the 37th month for Reading. It is probably unreasonable to believe that repeaters' scores would increase in the earlier times but decrease in the later times in 4 years. The cubic growth function showed that repeaters' scores increased faster in the earlier time, then increased slowly in the middle, and finally increased faster again in the later time. It seems that such a cubic growth trajectory is consistent with the general learning curve with plateau phase for many skills. Comparing the average quadratic and cubic growth trajectories with the observed means plots found that both models worked equally well for most of the time points, but the cubic model fit better with the observed data in the last few administration months for both Listening and Reading. Additional longitudinal data with more times of testing may provide stronger empirical evidence for the cubic growth model. However, the quadratic model was selected in this study based on the principle of parsimony in statistical modeling and the convenience in interpretation of growth parameters. Compared with the cubic model, the quadratic growth parameters are much easier to interpret for repeaters' score change patterns in the testing program. This is particularly true when examinees' background variables were included in the growth models.

The growth modeling results have important implications for the testing program. For the quality control of test performance, the repeaters' score change patterns can be used for the evaluation of the TOEIC Listening and Reading test scores from different perspectives. From a reliability perspective, the stability of repeaters' scores was a strong indicator of a high reliability of test scores across administrations, across forms, and over time. Reliability refers to the extent to which test scores are consistent across forms or occasions of testing. Therefore a testing program can evaluate test score reliability by examining form-to-form differences or differences in performance over time (ETS, 2014). The existence of many repeaters across administrations in the TOEIC Listening and Reading test provides empirical data to evaluate the consistency of test scores across forms and over time. The linear, quadratic, and cubic models in this study found very small monthly score increases. After taking account of the skill improvement due to learning or maturation, the test scores can be considered as consistent across forms and testing occasions.

From a validity perspective, the growth modeling results provided empirical evidence based on the relations of test scores to other variables. For validity evidence, the patterns of association between test scores and other variables should be consistent with theoretical expectations (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). This study included TIME and examinees' backgrounds as "other variables." The relations of TOEIC Listening and Reading test scores to these variables were consistent with related experience, theory, and previous studies. For example, (a) the repeaters' score growth trajectories (i.e., the relations of test scores to TIME) on the TOEIC were consistent with the repeaters' score change patterns found in other

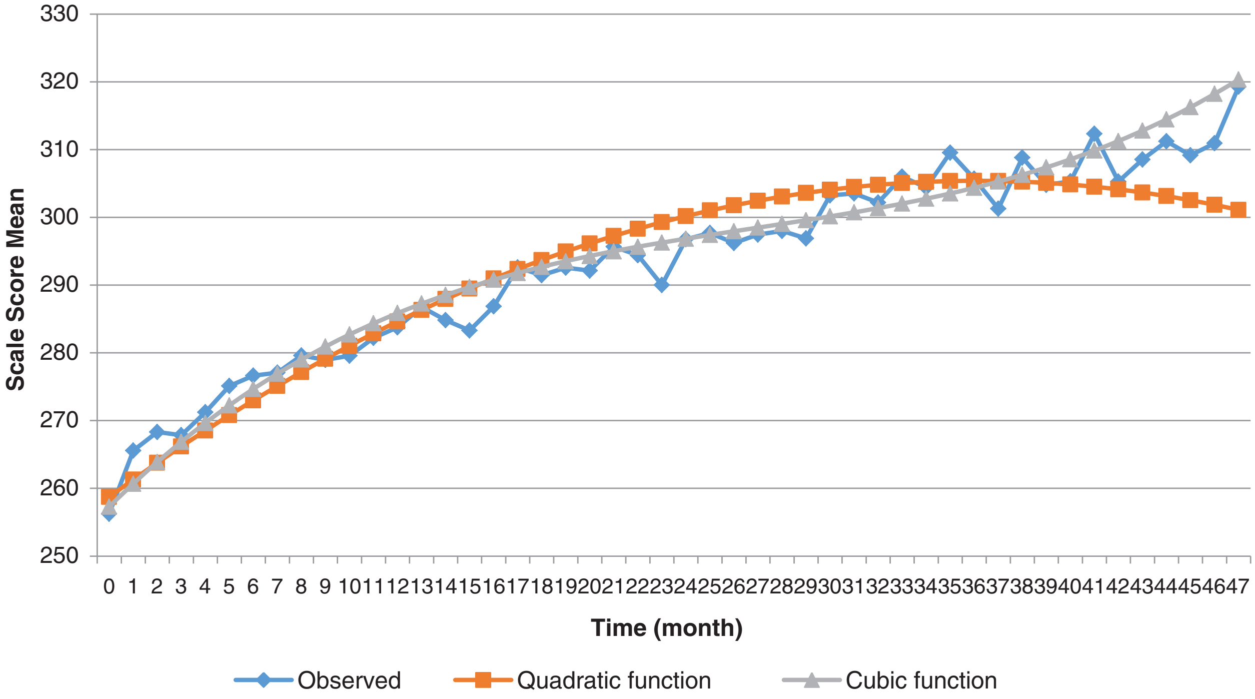

Figure 5 Growth functions and observed score means over time for Reading.

popular testing programs (e.g., TOEFL, GRE, and SAT), (b) the cubic growth trajectories (i.e., the relations of test scores to TIME) were consistent with the general learning curve for many skills, (c) the examinees' score increases with their daily English use time was consistent with language learning experience, and (d) the impact of examinees' educational level on their scores was consistent with English education experience.

From the test users' perspective, the gradual and stable score increase patterns supported the claim that TOEIC Listening and Reading test scores are suitable for measurement of progress in English proficiency over time. The small monthly growth rate reflected the stability of test scores, which supports the validity of the test scores for the intended use over time (von Davier, 2012). The increasing trend in test scores over time reflected repeaters' performance growth due to maturation and learning. Therefore test users can use TOEIC Listening and Reading test scores to evaluate English learning and training progress. In addition, the growth modeling results may help test takers or test users make decisions about retesting and learning strategies. For example, examinees with no previous test-taking experience tended to obtain lower scores than examinees with experience, but their scores increased more at the next testing. This may indicate that test takers can improve test scores by being familiar with the test and by retaking the test. Also, because examinees' scores increased with their daily English use time, one effective way to improve test scores is to use English more often in daily life. Furthermore, it is not unusual for English learners to make rapid progress at the beginning, then make slow or even little progress, and finally make apparent progress again if they keep learning. It may be helpful to know this learning curve for English learning and training.

The testing program can also use repeaters' growth modeling results to predict their performance on future administrations. The growth modeling results indicate that (a) the repeaters' fitted score means based on the models were consistent with their observed score means and (b) the growth parameters based on two different samples were very close to each other. This suggests that growth modeling is very promising for predicting repeaters' score means at the group level, which is consistent with findings from other studies (Wei, 2013; Wei & Qu, 2014). Therefore we can monitor test performance by comparing repeaters' observed and predicted score means. If repeaters' observed growth patterns are very different from the expected growth patterns, the testing program needs to investigate the inconsistency and find underlying reasons for it, such as population changes, scoring mistakes, or security breaches.

However, the testing program should be careful when making a judgment for an individual test taker. At the individual level, this study found that about 30% of score variance could be predicted based on the quadratic models. There were substantial individual variations in the growth trajectories, which is consistent with the findings of other studies (e.g., Wei & Morgan, 2016; Yang et al., 2011; Zhang, 2008). Therefore it is difficult to accurately predict individual repeaters' test scores. It is misleading to use the average growth trajectories to make a judgment about an individual's score change pattern.

To explore the individual variations in the group's growth trajectories, we have at least two ways to distinguish different growth patterns. One way is to use observed covariates to identify different patterns based on observed subgroups, as we did in this study. For example, adding the covariate test-taking experience helped us distinguish growth patterns for examinees without any test-taking experience and examinees with experience. More observed covariates can be included in the models to distinguish different patterns in future studies. The other way is to let the data distinguish growth patterns based on latent or underlying subgroups. Future studies can use latent class or mixture modeling methods to explore different latent score change patterns in the observed repeaters' data (e.g., Wei, 2016).

Conclusions

On the basis of the multilevel growth modeling analysis of the TOEIC Listening and Reading test scores of 19,855 examinees who had taken the test six times in 4 years, this study found that (a) examinees' scores increased with repeated testing; (b) examinees' score increase rates were higher in the early repetitions, then gradually dropped over time; (c) examinees without previous test-taking experience tended to have lower initial scores and higher initial increase rates, and their score increase rates tended to drop faster over time; (d) examinees' educational background had a significant relationship with their initial scores but had little association with their score increase rates; and (e) examinees' gender had some relationship to their initial scores and increase rates. The results suggest that multilevel growth modeling analysis can be used to evaluate test performance across administrations by exploring repeaters' score change patterns over time. Furthermore, growth modeling results support the reliability and validity of the TOEIC scores. The results also indicate that TOEIC scores can be used to evaluate English learning or training progress.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Washington, DC: Council of Chief State School Officers.

Educational Testing Service. (2013). *TOEIC user's guide*. Princeton, NJ: Author.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Kasim, R., & Raudenbush, S. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics, 20*, 93–116. https://doi.org/10.3102/10769986023002093

Kingston, N., & Turner N. (1984). *Analysis of score change patterns of examinees repeating the Graduate Record Examinations® General Test* (Research Report No. RR-84-22). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00062.x

Lee, Y.-H., & Haberman, S. J. (2013). Harmonic regression and scale stability. *Psychometrika, 78*, 815–829. https://doi.org/10.1007/s11336-013-9337-1

Lee, Y.-H., Liu, M., & von Davier, A. A. (2013). Detection of unusual test administrations using a linear mixed effects model. In R. Millsap, L. van der Ark, D. Bolt, & C. Woods (Eds.), *New developments in quantitative psychology: Proceedings of the 77th international meeting of the psychometric society* (Vol. 66, pp. 133–149). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-9348-8_9

Lee, Y.-H., & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika, 78*, 557–575. https://doi.org/10.1007/s11336-013-9317-5

Li, D., Li, S., & von Davier, A. A. (2011). Applying time-serious analysis to detect scale drift. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 327–346). New York, NY: Springer.

Nathan, J. S., & Camara, W. J. (1998). *Score change when retaking the SAT I: Reasoning Test* (Research Note No. RN-05). New York, NY: College Board.

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

Rock, D. R., & Werts, C. (1979). *An analysis of time-related increments and/or decrements for GRE repeaters across ability and sex groups* (GRE Board Research Report No. 77-9R). Princeton, NJ: Educational Testing Service.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001 von Davier, A. A. (2012). *The use of quality control and data mining techniques for monitoring scaled scores: An overview* (Research Report No. RR-12-20). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02302.x

Wei, Y. (2013). Monitoring TOEIC Listening and Reading test performance across administrations using examinees' background information. In D. E. Powers (Ed.), *The research foundation for TOEIC: A compendium of studies* (2nd ed., pp. 11.0–11.28). Princeton, NJ: Educational Testing Service.

Wei, Y. (2016, April). *Using growth mixture modeling to explore test takers' score change patterns*. Paper presented at the annual meeting of National Council on Measurement in Education, Washington, DC.

Wei, Y., & Morgan, R. (2016). *An evaluation of the single-group growth model (SGGM) as an alternative to common-item equating* (Research Report No. RR-16-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12087

Wei, Y., & Qu, Y. (2014). *Using multilevel analysis to monitor test performance across administrations* (Research Report No. RR-14-29). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12029

Wilson, K. M. (1987). *Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language* (Research Report No. RR-87-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00207.x

Yang, W., Bontya, A. M., & Moses, T. M. (2011). *Repeater effects on score equating for a graduate admissions exam* (Research Report No. RR-11-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02253.x

Zhang, Y. (2008). *Repeater analyses for TOEFL iBT*® (Research Memorandum No. RM-08-05). Princeton, NJ: Educational Testing Service.

Appendix

The unconditional means model is

$$Y_{ti} = \pi_{0i} + e_{ti}$$

$$\pi_{0i} = \beta_{00} + r_{0i},$$

where Y_{ti} is the test score of examinee i at administration time t; π_{0i} is the score mean of examinee i across administration times; e_{ti} is the residual or unique effect associated with examinee i at administration time t (i.e., within-person deviation) and is assumed to be normally distributed with $N\ (0,\sigma^2)$; β_{00} is the grand score mean (i.e., the average of all test takers' scores over time) of the population of test takers; and r_{0i} is the random effect associated with the examinee i (i.e., between-person deviation) and is assumed to be normally distributed with $N\ (0, \tau_{00})$.

The linear growth model is

$$Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{TIME}_{ti} + e_{ti},$$

$$\pi_{0i} = \beta_{00} + r_{0i},$$

$$\pi_{1i} = \beta_{10} + r_{1i},$$

where TIME_{ti} is the amount of time that had elapsed in months from the first time the examinee i took the test to administration time t; π_{1i} is the growth rate for examinee i over the 4 years of data collection and represents the expected change during a fixed unit of time (i.e., a month); π_{0i}, the intercept, is the initial status or the true score of examinee i at the first administration time (i.e., $\mathrm{TIME}_{ti} = 0$); e_{ti} is the deviation (i.e., residual) of examinee i at administration time t from his or her true linear growth trajectory; β_{00} and β_{10} are the mean intercept and mean linear growth rate that represent the mean growth trajectory of the population; and r_{0i} and r_{1i} are the deviations of examinee i's trajectory from the mean growth trajectory of the population in terms of initial status and linear growth rate, with a variance–covariance matrix:

$$\begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix},$$

where τ_{00} is the unconditional variance in the Level 1 intercepts, τ_{11} is the unconditional variance in the Level 1 growth rates, and τ_{01} or τ_{10} is the unconditional covariance between the Level 1 intercepts and linear growth rates.

The quadratic growth model is

$$Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{TIME}_{ti} + \pi_{2i}\mathrm{TIME}_{ti}^{2} + e_{ti},$$

$$\pi_{0i} = \beta_{00} + r_{0i},$$

$$\pi_{1i} = \beta_{10} + r_{1i},$$

$$\pi_{2i} = \beta_{20} + r_{2i},$$

where the linear component, π_{1i}, is the instantaneous growth rate for examinee i at the first administration time; the quadratic component, π_{2i}, is the acceleration in the growth trajectory; β_{00}, β_{10}, and β_{20} are the mean intercept, mean instantaneous growth rate, and mean acceleration of the population, respectively; and r_{0i}, r_{1i}, and r_{2i} are the deviations of examinee i's trajectory from the mean growth trajectory of the population in terms of initial status, instantaneous growth rate, and acceleration, respectively, with a variance–covariance matrix:

$$\begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{10} & \tau_{11} & \tau_{12} \\ \tau_{20} & \tau_{21} & \tau_{22} \end{pmatrix},$$

where the variance and covariance have similar interpretations in the linear growth model, with an addition of the acceleration component.

The conditional quadratic growth model is

$$Y_{ti} = \pi_{0i} + \pi_{1i}\mathrm{COV}_{ti} + \pi_{2i}\mathrm{TIME}_{ti} + \pi_{3i}\mathrm{TIME}_{ti}^2 + e_{ti},$$

$$\pi_{0i} = \beta_{00} + \beta_{01}X_i + r_{0i},$$

$$\pi_{1i} = \beta_{10} + r_{1i},$$

$$\pi_{2i} = \beta_{20} + \beta_{21}X_i + r_{2i},$$

$$\pi_{3i} = \beta_{30} + \beta_{31}X_i + r_{3i},$$

where the time-varying background covariate COV_{ti} was added in the Level 1 model and the person-level background variable X_i was added in the Level 2 models.

Suggested citation:

Wei, Y., & Low, A. (2017). *Monitoring score change patterns to support TOEIC® Listening and Reading test quality* (Research Report No. RR-17-54). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12186

Action Editor: Donald Powers

Reviewers: Lu Ru and Sooyeon Kim