Compendium Study

Monitoring *TOEIC*® Listening and Reading Test Performance Across Administrations Using Examinees' Background Information

Youhua Wei

September 2013

ETS TOEIC

Know English. Know Success.

For a testing program with many forms and administrations, test performance may fluctuate over time even though efforts are made to control the comparability of scores. Some contributing factors to test performance fluctuation include the evolution of test content, development in curriculum and training, population change, scale drift, rater drift, cumulative equating error, test difficulty shift, item exposure, and even operational mistakes. Among these contributing factors, some are not under the control of the testing company, such as population change, curriculum development, and training change; some can be controlled by the testing company, such as scale drift, rater drift, equating error, and test difficulty shift. Whether controllable or not, all these factors can impact the stability of test performance over time. Any unexpected fluctuation of test performance needs to be investigated, and any potential contributing factors need to be explored. Therefore, to ensure the quality of a testing program, it is important to understand and monitor the stability and fluctuation of test performance over administrations from different perspectives. As von Davier (2012) proposed, quality control in educational measurement is a formal systematic process that should be conducted not only within an individual administration but also across administrations during the life of a testing program. The across-administration quality control may include the evaluation of changes in examinees' background characteristics, subpopulation shift, seasonality of test performance, scale shift, test difficulty shift, and so on. Increasingly, studies have been conducted to address the quality control across administrations and different methods have been proposed or used for this purpose, such as time-series analysis (Li, Li, & von Davier, 2011), harmonic regression (Lee & Haberman, 2011), multivariate mixed weighted modeling (Luo, Lee, & von Davier, 2011), linear mixed effects modeling (Liu, Lee, & von Davier, 2012), Shewhart control charts (see a brief description in von Davier 2012), hidden Markov model (Lee & von Davier, in press; Visser, Raijmakers, & van der Maas, 2009), and multilevel analysis (Wei & Qu, 2012).

However, most across-administration quality control studies have focused on methodologies and techniques (von Davier, 2012) or equating effects (e.g., Haberman & Dorans, 2011; Haberman, Guo, Liu, & Dorans, 2008). Very few studies have explored how to use the relationship between examinees' characteristics and their test performance to monitor and control the quality of the test performance across administrations. It is not unusual for a testing program to collect examinees' background information during the registration or administration of the test. Both empirical research (e.g., Liu et al., 2012; Luo et al., 2011; Wei & Qu, 2012) and operational experience suggest that relationships exist between examinees' test scores and their background. Some studies have been conducted to explore the potential of using examinees' background information for improving equating accuracy. Unfortunately, the general conclusion is that examinees' background information does not provide much additional information for equating (Kolen, 1990; Paek, Liu, & Oh, 2010), and its use has not been recommended to adjust group ability difference for equating purposes (Liao & Livingston, 2012). Although examinees' background information does not help much with equating per se, it has the potential for monitoring reported test scores after equating has been conducted. As Allalouf (2007) suggested, exploring the statistical relationship between examinees' background information and their scores should be part of the quality control procedure for a testing program. The relationship can then be used to understand and monitor test scores. Three studies (Liu et al., 2012; Luo et al., 2011; Wei & Qu, 2012) found statistically significant relationships of examinees'

background variables with their test performance. However, the small number of administrations or the short test lengths used in those studies make it difficult to fully identify close relationships and establish powerful prediction models. It was not very clear how practically or psychometrically significant the examinees' background association with their scores was, or more importantly, how well the resulting prediction models could be used to understand and monitor test performance for future administrations.

In the operational work for a testing program, different methods and procedures can be used to explore the statistical relationship between examinees' background and their scores. For example, scatter plots and correlation coefficients are simple and straightforward methods to find the bivariate relationship; regression can be used to predict examinees' scores from one or multiple background variables. With individual examinees' background variables appropriately coded and administrations' background variables carefully created, these methods can be used simultaneously or separately at the examinee level and at the administration level. For a testing program with many administrations, test data have a two-level hierarchical structure, with examinees at Level 1 and administrations at Level 2. Given that the statistical relationships between examinees' background and test scores may vary across administrations, multilevel analysis (i.e., hierarchical linear modeling, Snijders & Bosker, 1999; Raudenbush & Bryk, 2002) has potential in exploring both the random and fixed relationships among variables at different levels (Wei & Qu, 2012).

A previous multilevel analysis study (Wei & Qu, 2012) using English speaking performance assessment data found that examinees' background information has potential for predicting test scores especially at the administration level. Both fixed and random effects should be considered while evaluating test performance across administrations. This previous study was based on a very short English speaking test with only 13 constructed response items and no equating was conducted. Therefore, the potential relationships between examinees' test scores and their background may not have been fully identified.

The current study used multilevel analysis to explore the relationships between examinees' background and their scale scores on the *TOEIC*® Listening and Reading test. Specifically, the study addressed the following questions:

1. How strong are the relationships between examinees' test scale scores and their background information at the individual level? Are those relationships consistent, or do they vary across different administrations?
2. How strong are the relationships between examinees' scale score means and their background information at the administration level?
3. Can the relationships found in Questions 1 and 2 be used to monitor examinees' test performance at the examinee level and at the administration level?

Methodology

Data

The data were collected from the TOEIC Listening and Reading test, which is designed to evaluate examinees' English listening and reading skills in two sections. Each section comprises 100 multiple-choice items. Raw scores range from 0 to 100, and the reported scale scores range from 5 to 495 by increments of 5. Equating is conducted separately for the two sections to obtain examinees' scale scores.

A background questionnaire is used to collect information on examinees' educational and work-related background, English-language experience, and test-taking experience. Specifically, 14 questions in the questionnaire offer different options: Five questions cover examinees' educational and/or work-related background (e.g., "Choose either the level of education in which you are currently enrolled or the highest level that you have completed?"), seven questions are about the examinees' English-language experience (e.g., "How many years have you spent studying English?"), and two questions address examinees' experience in taking the test ("What is your main purpose for taking today's test?"). The examinees' responses to these questions were coded for analyses.

The study was based only on the data of those examinees who had taken the TOEIC Listening and Reading test for the first time. The inclusion of repeaters' data might have violated the assumption of independence of observations in different Level 2 units for multilevel analysis. Table 1 shows the summary statistics of the test scores at both examinee and administration levels. The data include 1,499,313 examinees' TOEIC Listening and Reading scale scores and their background information collected from 71 administrations of the test in Korea within 6 years. With each form being used in each administration, the sample sizes range from 5,295 to 40,615, with an average of 21,117; the TOEIC Listening scale score means range from 255.60 to 322.40 with an average of 283.77; the TOEIC Reading scale score means range from 202.30 to 258.10 with an average of 228.92.

Table 1

Summary Statistics of Test Scores at Level 1 and Level 2

Data level	TOEIC section score/statistic	*N*	Mean	SD	Minimum	Maximum
Level 1 (examinee)	Listening score	1,499,313	285.01	89.00	10.00	495.00
	Reading score	1,499,313	230.33	97.44	5.00	495.00
Level 2 (administration)	Listening mean	71	283.77	15.77	255.60	322.40
	Listening SD	71	87.98	3.52	78.00	95.50
	Reading mean	71	228.92	15.51	202.30	258.10
	Reading SD	71	96.41	4.36	87.20	111.50
	Sample size	71	21,117	8,408	5,295	40,615

Procedure and Analyses

The analyses were conducted separately for TOEIC Listening and Reading sections because the sections were designed to measure two different constructs and equatings were conducted separately.

Data preparation. The data of 1,499,313 examinees' test scores and their responses to the 14 background questions were reorganized at two levels. At Level 1, the scale scores of all individual examinees were used as the dependent variables. The coding of examinees' background responses as predictor variables was based on each background question's original response options. However, for the sake of parsimony in statistical modeling, some response options of the questions were combined if different subgroups based on the options had consistently similar test performance across most administrations. At Level 2, the scale score means of each administration were used as the dependent variables. The predictors for administration means were group composition variables, which were defined as the percentages of one or more combined subgroup(s) based on examinees' responses to background questions in specific administrations. The following section describes in detail how the background variables were coded and selected at both levels.

Preliminary analyses and variable selection. To explore the relationships between examinees' test scores and their background information, it was important to carefully select important background questions and code examinees' responses in an informative and simple way. Some preliminary analyses were conducted for this purpose.

At Level 1, the examinees' background variables based on the questionnaire were categorical, so bivariate correlations and scatter plots were not appropriate to explore their relationships with test scores. Instead, for each background question, the scale score means of subgroups based on different options were plotted and compared across administrations. If no clear and consistent patterns of score means were evident between different subgroups across administrations, the background question was not selected for further analyses. If the mean differences between subgroups were consistent across administrations, those subgroups were coded as different

subgroups and given different values. However, if the mean differences between subgroups were trivial across administrations, those subgroups were coded as the same subgroup and given the same value. For example, for the background question about test-taking purpose, the subgroup selecting "for program evaluation," and the subgroup selecting "for course graduation" had consistently different scale score means in most administrations, so they were coded into different subgroups and given different values. In this study, the subgroup selecting for job application and the subgroup selecting "for learning evaluation" had very similar scale score means across all administrations, so they were combined together and coded into the same value. It was not unusual that some examinees did not respond to some background questions, so for each question, there was a special subgroup with missing information. However, it was found that compared with other subgroups, the special subgroup with missing information might have consistent performance across administrations. Therefore, these special subgroups were not automatically excluded from the data. Instead, they were coded based on their consistent performance across administrations. For example, for the background question about test-taking purpose, the examinees with missing information had very similar means as the examinees selecting "for job application" and "for learning evaluation," so they were coded into the same value.

Using this rationale, eight background questions were selected and the responses to these questions were coded for further analyses (see Table 2). Of the eight selected background questions in Table 2, four were coded as nominal variables (i.e., educational level, current occupation, overseas living purpose, and test-taking purpose) because only qualitative differences were found among their values. Another four questions were coded as ordinal variables (i.e., English study time, daily English use time, English communication difficulty, and overseas English experience) because an order was present in their values. Some subgroups were combined to form a new single subgroup due to their similar performance across administrations. For example, for the background question about current occupation, the examinees who chose "unemployed" and the examinees with missing information were combined into one single subgroup. Table 2 also shows the average subgroups' percentages across all administrations.

Table 2

Level 1 Variables and Codes for Both the TOEIC Listening and Reading Sections

Background question	Option	Code	Subgroup percentage	Variable name
Education level	Vocational/technical school	0 (reference)	11.94	SQ1R1
	Community/junior college			
	Missing	1	88.06	
	High school and below			
	Undergraduate			
	Graduate			
	Language institute			
Current occupation	Full-time employed	(0, 0, 0) (Reference)	8.28	
	Part-time employed	(1, 0, 0)	2.71	SQ3R1
	Unemployed	(0, 1, 0)	9.37	SQ3R2
	Missing			
	Full-time student	(0, 0, 1)	79.64	SQ3R3
English study time	≤ 4 years	1	10.87	SQ6
	4–6 years	2	10.48	
	6–10 years	3	48.98	
	Missing			
	>10 years	4	29.67	
Daily English use time	None	1	64.55	SQ8
	Missing			
	1–10%			
	11–20%	2	22.28	
	21–50%	3	10.27	
	51–100%	4	2.9	
English communication difficulty	Almost never	1	2.42	SQ10
	Seldom	2	5.3	
	Sometimes	3	24.37	
	Missing	4	42.58	
	Frequently			
	Almost always	5	25.33	
Overseas English experience	Missing	1	76.4	SQ11
	None			
	<6 months	2	13.08	
	6–12 months	3	5.05	
	1–2 years	4	2.26	
	>2 years	5	3.2	

Background question	Option	Code	Subgroup percentage	Variable name
Overseas living purpose	Missing	(0, 0, 0) (Reference)	68.45	
	Travel	(1, 0, 0)	18.14	SQ12R1
	Work			
	Other			
	Study in English program	(0, 1, 0)	8.32	SQ12R2
	Study in non-English program	(0, 0, 1)	4.82	SQ12R3
Test-taking purpose	Promotion	(0, 0, 0) (Reference)	2.34	
	Missing	(1, 0, 0)	71.96	SQ14R1
	Job application			
	Learning evaluation			
	Course graduation	(0, 1, 0)	22.06	SQ14R2
	Program evaluation	(0, 0, 1)	3.63	SQ14R3

At Level 2, the group composition variable was defined as the percentage of the subgroup(s) based on examinees' responses to a specific background question in the administration. The bivariate scatter plots and correlations between group composition variables and test score means were used to explore their relationships. Based on the preliminary analyses, 13 group composition variables were defined and coded as predictors for administrations' score means (see Table 3).

Table 3

Level 2 Variables and Codes for Both the TOEIC Listening and Reading Sections

Background question	Variable name	Definition/coding	Mean	SD	Min	Max
Education level	SQ0106	% with community/junior college level	11.99	2.29	7.32	17.04
Major	SQ0205	% with health major	5.22	1.22	3.17	8.27
Current occupation	SQ0301	% of full-time employed	9.17	2.96	4.37	18.99
English study time	SQ0602	% studying English 4-6 years (for TOEIC Listening mean scores)	10.70	1.04	8.61	13.40
	SQ0601	% studying English ≤4 years (for TOEIC Reading mean scores)	11.11	1.22	8.59	13.58
English skills emphasized	SQ0705	% emphasizing listening/speaking	35.80	1.35	33.28	38.94
Daily English use time	SQ0812	% using English less than 10% daily time	58.54	1.86	54.00	62.69
English skills most used	SQ09CO	% using writing, listening/speaking, and reading/writing (for TOEIC Listening mean scores)	35.00	1.04	33.07	37.35
	SQ0902	% using reading (for TOEIC Reading mean scores)	21.42	1.10	19.17	23.59
Overseas English experience	SQ1145	% with over 1 year overseas English experience	5.75	1.52	3.60	10.45
Overseas living purpose	SQ1201	% living overseas for study (for TOEIC Listening mean scores)	5.07	1.57	2.78	10.03
	SQ1212	% living overseas for study or in English program (for TOEIC Reading mean scores)	13.42	1.73	10.64	18.15
Test-taking purpose	SQ1402	% taking test for job promotion	2.59	0.86	1.35	5.08

Table 3 shows how the 13 group composition variables at the administration level were created from the 13 background questions. The definition and coding of these Level 2 predictors were mainly based on their relationships with administrations' score means. Note that the same background question might be coded in different ways for the listening and reading sections on the *TOEIC*® Listening and Reading test due to different degrees of relationships found in the bivariate plots and correlations. For example, the group composition variable based on the background question about English study time was defined and coded as "% studying English 4–6 years" for TOEIC Listening score means but as "% studying English ≤ 4 years" for TOEIC Reading score means. Similarly, group composition variables were defined and coded differently for the background questions regarding English skills most used and overseas living purpose. Therefore, the study used 10 group composition variables for TOEIC Listening score means and 10 variables for TOEIC Reading score means. Table 3 also shows the means, standard deviations, minimums, and maximums of each group composition variable across administrations at Level 2.

To find the best prediction models, all possible subsets regression analyses based on *R* square were conducted to explore the best background predictors for test performance at both the examinee and the administration levels. The best models identified would be used to explore the best models in the following multilevel analysis.

Multilevel analysis. Two-level hierarchical linear modeling was used to investigate the relation of examinees' background to their test scores across administrations, with examinees at Level 1 and administrations at Level 2. Based on the preliminary analysis results, different models were explored and results were evaluated in terms of prediction of test scores based on examinee's background information. Specifically, three models were used in the study.

The first model is one-way analysis of variance (ANOVA) model with random effect,

$$\text{Level 1: } Y_{ij} = \beta_{0j} + r_{ij} \text{ ;}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j},$$

where Y_{ij} is the test score of examinee i on administration j; β_{0j} is the score mean of examinees on administration j; r_{ij} is the residual or unique effect associated with examinee i on administration j and is typically assumed to be normally distributed with N (0, σ^2); γ_{00} is grand score mean (i.e., the average of administration score means) in the population of administrations; u_{0j} is the random effect associated with administration j and is typically assumed to be normally distributed with N (0, τ_{00}).

The second model is regression with means-as-outcomes,

$$\text{Level 1: } Y_{ij} = \beta_{0j} + r_{ij} \text{ ;}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01} G_j + u_{0j} \text{ ,}$$

where G_j is the Level 2 predictor or group composition variable for score mean on administration j; γ_{01} is the slope in regression of β_{0j} on predictor G_j; γ_{00} is the grand score mean conditioned on the predictor G_j; u_{0j} is the random effect associated with administration j conditioned on the predictor G_j ,with a normal distribution N (0, τ_{00}).

The third model is random-coefficient model,

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j} B_{1ij} + r_{ij} \text{ ;}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \beta_{1j} = \gamma_{10} + u_{1j} \text{ ,}$$

where B_{1ij} is the predictor or examinee background variable at Level 1; β_{0j}, β_{1j} are intercept and slope in regression of Y_{ij} on Level 1 predictor; r_{ij} is the residual conditioned on Level 1 predictor; γ_{00} and γ_{10} are the grand mean and average slope in the population of administrations; u_{0j} and u_{1j} are the intercept's and slope's random effects associated with administration j, 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 slopes, and τ_{01} or τ_{10} is the unconditional covariance between the Level 1 intercepts and slopes.

The multilevel analyses started with an ANOVA model with an evaluation of the Level 1 variance. Then a regression with means-as-outcomes model was used to evaluate the relationship between administration means and group composition at Level 2. The random-coefficient model was used to explore the relationship between examinees' test scores and their background at Level 1, and the variance component estimates were used to check the consistency of the relationship across administrations. For convenience in operational use and interpretation, the analyses started with one level with the other level being held aside. Each background variable was first separately used in the model and then more predictors were included at the same level until the best prediction model was identified. However, due to iteration time in computation and possible difficulty in interpretation, efforts were made to avoid including too many predictors in one model unless additional predictors could significantly improve predictive power and accuracy. For the metric of predictors at Level 1, the four ordinal background variables used their natural scale and the four nominal variables were dummy-coded with one subgroup as the reference group; at Level 2, the 13 group composition variables used their natural scale. These metrics should be considered while interpreting results from the analyses. HLM 6.06 (Raudenbush, Bryk, & Congdon, 2000) was used for all multilevel analyses in this study and full maximum likelihood estimation method was selected for all models.

Model validation. For the strong prediction models, examinees' background data from other later administrations not included in the modeling were used to predict their test performance (e.g., group means in those administrations). The predicted scale scores or scale score means were then compared with those produced from operational equating and scoring. The results were used to evaluate how accurate and stable the models performed in predicting test performance, especially at the administration level.

Results

In this section, I first summarize the one-way ANOVA results, which provide baseline information for further analyses. Then I explore the prediction of scale score means at the administration level by using a regression with means-as-outcomes model and the prediction of scale scores at the examinee level by using a random-coefficient regression model. I close the section by applying the strong prediction models identified in the analyses to the new operational data and evaluating their stability and accuracy.

Estimating Variance Components: One-Way ANOVA With Random Effect

As the simplest model, the one-way ANOVA model provides preliminary results about the variation of test scores within and between administrations. It also provides reliability estimates of observed administration means (i.e., sample means) for the true means of the populations on those administrations. For the purpose of this study, the one-way ANOVA model provided the Level 1 and Level 2 base variances, which were used in further analyses to evaluate how strongly the examinees' background information predicted their test performance.

Listening. Based on the results from the one-way ANOVA model with random effect and with homogeneity assumption of Level 1 variance σ^2 (see Table 4 for the detailed results), the grand mean of test scores across administrations was $\hat{\gamma}_{00}$ = 283.77, with a standard error of 1.86. So the 95% confidence interval for the grand mean was 283.77 ± 1.96 * 1.86 = (280.12, 287.42). Although the reliability of the sample mean as an estimate of the true mean might vary across administrations due to different sample sizes, an overall reliability estimate of the observed sample means $\hat{\beta}_{0j}$ was $\hat{\lambda}$ = 0.998. Therefore, the grand mean estimate appeared to be very precise and the sample means tended to be very reliable estimates of the true score means.

Table 4

Results From ANOVA Model With Homogeneous σ^2 for the TOEIC Listening Section $Y_{ij} = \beta_{0j} + r_{ij}$; $\beta_{0j} = \gamma_{00} + u_{0j}$

Effect		Coefficient	Variance component	SE	SD	T-ratio	χ^2	*df*	*p*
Fixed	Average admin mean γ_{00}	283.77		1.86		152.76		70	<.001
Random	Admin mean u_{0j}		248.06		15.75		46545.42	70	<.001
	Level 1 effect r_{ij}	7684.97		87.66					

Note. Random Level 1 coefficient β_{0j}'s reliability estimate = 0.998.

For the variance components, Level 1 variance was $\hat{\sigma}^2$ = 7684.97 and Level 2 variance was $\hat{\tau}_{00}$ = 248.06. The intraclass correlation $\hat{\rho} = \hat{\tau}_{00} / (\hat{\tau}_{00} + \hat{\sigma}^2)$ = 248.06 / (248.06 + 7684.97) = 0.0313 indicates that 3.13% of the variance in the test scores was between administrations. Therefore, most score variation came from within administrations. The low intraclass correlation also suggests a lower degree of dependence of examinees' scores within administrations. However, the between-administration variance of 248.06 was still significantly larger than 0, with $\chi^2 = 46545.41, df = 70, p < 0.001$. The 95% confidence interval of the administration means falls within the range 283.77 ± 1.96 * $\sqrt{248.06}$ = (252.90, 314.64). Based on the criterion used in Shewhart 3-sigma control charts, the lower control limit (LCL) was 236.52 and the upper control

limit (UCL) was 331.02. Therefore, psychometrically, the scale score means fluctuated across administrations and the between-administration variance should not be ignored in analyses.

Reading. The one-way ANOVA model with random effect produced the following results for TOEIC Reading scores (see Table 5 for the detail): $\hat{\gamma}_{00} = 228.93$ with standard error of 1.83, 95% confidence interval for $\hat{\gamma}_{00} = (225.34, 232.52)$, $\hat{\lambda} = 0.998$, $\hat{\sigma}^2 = 9257.49$, $\hat{\tau}_{00} = 240.12$, $\hat{\rho} = \hat{\tau}_{00} / (\hat{\tau}_{00} + \hat{\sigma}^2) = 0.0253 = 2.53\%$, 95% confidence interval for $\hat{\beta}_{0j} = (198.55, 259.31)$, and an LCL and UCL of 182.43 and 275.43, respectively.

Table 5

Results From ANOVA Model With Homogeneous σ^2 for the TOEIC Reading Section

$Y_{ij} = \beta_{0j} + r_{ij}$; $\beta_{0j} = \gamma_{00} + u_{0j}$

Effect		Coefficient	Variance component	SE	SD	T-ratio	χ^2	*df*	*p*
Fixed	Average admin mean γ_{00}	228.93		1.83		125.23		70	<.001
Random	Admin mean u_{0j}		240.12		15.50		38812.36	70	<.001
	Level 1 effect r_{ij}		9257.49		96.22				

Note. Random level-1 coefficient β_{0j}'s reliability estimate = 0.998.

Therefore, for both the TOEIC Listening and Reading sections, most score variation came from within administrations. This is consistent with the finding from a multilevel study of a speaking test (Wei & Qu, 2012) that 3.9% of the variance in the test scores was between administrations. However, psychometrically, the between-administration variance was not trivial and the scale score means fluctuated substantially across administrations.

These results were based on the ANOVA model with the assumption of homogeneity of Level 1 variance σ^2. The likelihood ratio test suggests that σ^2 was not homogeneous across administrations for both TOEIC Listening and Reading scores. However, given that the estimation of fixed effects and their standard errors was robust to the violation of this assumption (Kasim & Raudenbush, 1998) and that a general estimate of σ^2 was needed to estimate variance explained by Level 1 predictors, the results from ANOVA model with the assumption of homogeneity of σ^2 were used for this study (i.e., $\hat{\sigma}^2 = 7684.97$ and $\hat{\tau}_{00} = 248.06$ for TOEIC Listening, and $\hat{\sigma}^2 = 9257.49$ and $\hat{\tau}_{00} = 240.12$ for TOEIC Reading).

Predicting Test Performance at the Administration Level: Regression With Means as Outcomes

A means-as-outcomes regression model was used to explore the relationship between examinees' test performance and their background information at the administration level. This type of model

was first used separately for each of the selected group composition variables then used for the combination of those variables to predict administration score means.

Listening. Table 6 summarizes the results from each of the 10 single-predictor models and from the best combined-predictor model for TOEIC Listening score means. The best prediction model was selected based on the amount of variance explained and the significance of prediction coefficients. The $\hat{\gamma}_{00}$ column in Table 6 shows the intercepts or the grand means conditioned on predictors (i.e., when the predictors had a value of zero). Because I used group composition variables at their original metrics and no predictors' values were equal to zero in the data, it is hard to explain the exact meanings of those grand mean estimates. The first focus is on each predictor's coefficient and examines how each group composition variable was related to groups' score means. From the $\hat{\gamma}_{01}$ column in the table, each group composition variable had a significant relationship with group means. For example, when the percentage of examinees with community/junior college education level in the group increased by 1, the group's score mean decreased by 4.69 scale score points; when the percentage of examinees with health majors in the group increased by 1, the group's mean increased by 8.25 score points; when the percentage of full-time employed examinees in the group increased by 1, the group mean decreased by 2.91 points.

Table 6

Results From Means-as-Outcome Regression Models for the TOEIC Listening Section

$Y_{ij} = \beta_{0j} + r_{ij}$, $\beta_{0j} = \gamma_{00} + \gamma_{01} Group_j + u_{0j}$

Level 2 predictor	Variable name	$\hat{\gamma}_{00}$	$\hat{\gamma}_{01}$ *(P)*	$\hat{\tau}_{00}$	Variance explained %
Education level	SQ0106	339.99	-4.69 (<.001)	135.09	45.54
Major	SQ0205	240.77	8.25 (<.001)	148.84	40.00
Current occupation	SQ0301	310.47	-2.91 (<.001)	176.36	28.90
English study time	SQ0602	394.73	-10.37 (<.001)	133.43	46.21
English skills emphasized	SQ0705	563.96	-7.83 (<.001)	138.59	44.13
Daily English use time	SQ0812	565.12	-4.81 (<.001)	170.92	31.10
English skills most used	SQ09CO	687.99	-11.55 (<.001)	105.90	57.31
Overseas English experience	SQ1145	257.16	4.63 (.001)	201.23	18.88
Overseas living purpose	SQ1201	256.83	5.31 (<.001)	180.79	27.12
Test-taking purpose	SQ1402	305.29	-8.32 (<.001)	199.51	19.57
Education level	SQ0106	462.86	$\hat{\gamma}_{01}$ = -2.27 (<.001)	38.40	84.52
Major	SQ0205	462.86	$\hat{\gamma}_{02}$ = 2.60 (.002)	38.40	84.52
English study time	SQ0602	462.86	$\hat{\gamma}_{03}$ = -4.48 (<.001)	38.40	84.52
English skills most used	SQ09CO	462.86	$\hat{\gamma}_{04}$ = -3.78 (.001)	38.40	84.52
Overseas English experience	SQ1145	462.86	$\hat{\gamma}_{05}$ = 2.56 (<.001)	38.40	84.52

Note. $\hat{\sigma}$ = 87.66.

To evaluate the predictive power of each group composition variable on a group's score means, one can estimate the proportion of means' variance explained by that group composition variable using

$$\frac{\hat{\tau}_{00}(randomANOVA) - \hat{\tau}_{00}(groupcompostion)}{\hat{\tau}_{00}(randomANOVA)}.$$

Based on the one-way ANOVA with random effect, $\hat{\tau}_{00}$ = 248.06. The $\hat{\tau}_{00}$ column in Table 6 shows $\hat{\tau}_{00}$ estimates based on different regression models with means-as-outcomes. The last column in the table shows the percents of means' variance explained by different group composition variables. For example, for the group composition variable based on education level, the explained means' variance percentage was (248.06 – 135.09) / 248.06 = 45.54%. Table 6 shows that a single group composition variable could predict from 19% to 57% of the variance of score means. The best single predictor was the group composition variable based on English skills most used (i.e., the percentage of examinees who most often used writing, listening and speaking, and reading and writing skills).

The bottom row in Table 6 shows the results from the best combined-predictor model, which was identified based on variance explained and statistical significance of each predictor in the model. Based on this model, the best combined group composition variables for TOEIC Listening score means were percentages of (a) examinees with community/junior college education, (b) examinees with health majors, (c) examinees who have learned English for 4 to 6 years, (d) examinees who most often used writing, listening and speaking, and reading and writing skills, and (e) examinees with more than 1 year of overseas English experience. These five group composition variables could explain 85% variance of administrations' score means (i.e., *R* square = 0.85), and the root mean squared error (RMSE) was $\sqrt{38.40}$ = 6.20.

Reading. Table 7 shows the basic results from each of the 10 single-predictor models and from the best combined-predictor model for TOEIC Reading score means. As shown in the TOEIC Listening section, each group composition variable had a significant relationship with the group means. For example, when the percentage of examinees with community/junior college education level in the group increased by 1, the group's score mean decreased by 5.30 scale score points; when the percentage of examinees with health majors in the group increased by 1, the group's mean increased by 6.87 score points; when the percentage of full-time employed examinees in the group increased by 1, the group mean decreased by 2.68 points. One single group composition variable could predict from 20% to 61% of the variance of score means. The best single predictor was the percentage of examinees with community/junior college as their highest education level.

Table 7

Results From Means-as-Outcome Regression Models for the TOEIC Reading Section

$$Y_{ij} = \beta_{0j} + r_{ij}; \beta_{0j} = \gamma_{00} + \gamma_{01} Group_j + u_{0j}$$

Level 2 predictor	Variable name	$\hat{\gamma}_{00}$	$\hat{\gamma}_{01}$ *(P)*	$\hat{\tau}_{00}$	Variance explained %
Education level	SQ0106	292.51	-5.30 (<.001)	94.46	60.66
Major	SQ0205	193.07	6.87 (<.001)	172.16	28.30
Current occupation	SQ0301	253.53	-2.68 (<.001)	179.64	25.19
English study time	SQ0601	312.70	-7.54 (<.001)	158.26	34.09
English skills emphasized	SQ0705	492.07	-7.35 (<.001)	143.94	40.05
Daily English use time	SQ0812	453.86	-3.84 (<.001)	192.05	20.02
English skills most used	SQ0902	29.83	9.29 (<.001)	138.01	42.52
Overseas English experience	SQ1145	201.84	4.71 (.001)	191.44	20.27
Overseas living purpose	SQ1212	162.47	4.95 (<.001)	168.75	29.72
Test-taking purpose	SQ1402	250.63	-8.39 (<.001)	190.56	20.64
Education level	SQ0106	247.50	$\hat{\gamma}_{01}$ = -3.44 (<.001)	37.43	84.41
Major	SQ0205	247.50	$\hat{\gamma}_{02}$ =3.08 (<.001)	37.43	84.41
Daily English use time	SQ0812	247.50	$\hat{\gamma}_{03}$ = -1.16 (.02)	37.43	84.41
English skills most used	SQ0902	247.50	$\hat{\gamma}_{04}$ =2.68 (.007)	37.43	84.41
Overseas English experience	SQ1145	247.50	$\hat{\gamma}_{05}$ =3.04 (<.001)	37.43	84.41

Note. $\hat{\sigma}$ = 96.22.

Based on the best prediction model results shown on the bottom of Table 7, the best combined group composition variables for TOEIC Reading score means were percentages of (a) examinees with community/junior college education level, (b) examinees with health majors, (c) examinees using English less than 10% of the time in their daily life, (d) examinees who most often used the reading skill, and (e) examinees with more than 1 year of overseas English experience. These five group composition variables could together explain 84% variance of TOEIC Reading score means (i.e., *R* square = 0.84) and the RMSE was $\sqrt{37.43}$ = 6.12.

Therefore, the results from the means-as-outcomes regression models suggest that there is a very strong relationship between group composition variables and group means for both TOEIC Listening and Reading sections. Using five group composition variables as predictors, around 85% of score means' variance could be explained for both sections. In other words, the correlation coefficient between score means and group composition was as high as 0.92 for both the TOEIC Listening and Reading sections. The prediction error was 6 score points for both sections.

When the interactions between the five group composition variables were added in the best prediction models, the *R* square increased by 0.04 for the TOEIC Listening section and by 0.03 for the TOEIC Reading section and the RMSE decreased by 0.47 for the TOEIC Listening section and by 0.13 for the TOEIC Reading section. However, most prediction coefficients in the models became statistically nonsignificant for both sections. Therefore, adding interactions in the models did not improve the prediction.

To test for the possibility effect of seasonality (i.e., higher test score means in some months than in others), a new dummy-coded administration variable, seasonality, was added to the best Level 2 models to check its impact. The *R* square difference between the new and original models (i.e., 0.0003 for the TOEIC Listening section and 0.0021 for the TOEIC Reading section) was very small, and the prediction coefficient of seasonality in the new models was not significant (i.e., $p = 0.70$ for the TOEIC Listening section and $p = 0.35$ for the TOEIC Reading section). Therefore, adding the administration variable as seasonality predictor in the models did not improve the predictive power. The group composition variables seem to have already accounted for any apparent seasonality of scale score means across administrations.

A regular regression model with score means as the dependent variable and group composition as the independent variables (i.e., using only Level 2 data) was also conducted and the same best model was identified for both the TOEIC Listening and Reading sections. The regular regression model was very similar to the multilevel regression with means-as-outcomes model in terms of predictors, regression coefficient estimates, *R* square and adjusted *R* square, and RMSE, with the latter's prediction error being slightly lower.

Predicting Test Performance at the Examinee Level: Random-Coefficient Regression

A random-coefficient regression model was used to explore the relationship between examinees' test performance and their background information at the individual examinee level. This type of model was first used separately for each of the selected Level 1 background variables and then used for the combination of those variables to predict examinees' scores.

Listening. Table 8 summarizes the results from each of the eight single-predictor models and from the best combined-predictor model for TOEIC Listening scores. The $\hat{\gamma}_{00}$ and $\hat{\gamma}_{10}$ columns in Table 8 show the average intercepts and slopes estimated for the Level 1 intercepts and slopes for predicting examinees' scores on their background variables. To check the variability of the parameter estimates, the standard deviations of the average intercepts and slopes were also included in the table. All the intercepts and slopes and their variances were statistically significant with $p < 0.001$ (not shown in the table).

Table 8

Results From Random-Coefficient Regression Models for the TOEIC Listening Section

$Y_{ij} = \beta_{0j} + \beta_{1j}B_{1ij} + r_{ij}$, $\beta_{0j} = \gamma_{00} + u_{0j}$, $\beta_{1j} = \gamma_{10} + u_{1j}$

Level 1 predictor	Variable name	$\hat{\gamma}_{00}$ (SD)	$\hat{\gamma}_{10}$ (SD)	$\hat{\sigma}^2$	Variance explained %
Education level	SQ1R1	234.15 (11.98)	56.50 (6.26)	7357.89	4.26
Current occupation	SQ3R1	253.71 (8.74)	8.65 (6.13)	7562.49	1.59
	SQ3R2		26.79 (7.72)		
	SQ3R3		34.60 (12.19)		
English study time	SQ6	225.52 (14.06)	19.69 (1.10)	7361.38	4.21
Daily English use time	SQ8	256.43 (16.82)	18.00 (3.65)	7490.76	2.53
English communication difficulty	SQ10	394.14 (33.68)	-28.96 (6.19)	6880.12	10.47
Overseas English experience	SQ11	229.73 (18.96)	37.47 (3.28)	6505.11	15.35
Overseas living purpose	SQ12R1	267.76 (15.97)	24.63 (2.90)	6827.29	11.16
	SQ12R2		64.03 (8.13)		
	SQ12R3		117.35 (11.57)		
Test-taking purpose	SQ14R1	241.44 (12.35)	38.58 (10.75)	7545.90	1.81
	SQ14R2		55.70 (16.37)		
	SQ14R3		69.26 (13.48)		
English study time	SQ6	210.05 (25.77)	16.43 (1.14)	5594.81	27.20
English communication difficulty	SQ10	210.05 (25.77)	-17.75 (4.45)	5594.81	27.20
Overseas English experience	SQ11	210.05 (25.77)	27.55 (3.71)	5594.81	27.20
Education level	SQ1R1	210.05 (25.77)	41.03 (4.36)	5594.81	27.20
Current occupation	SQ3R3	210.05 (25.77)	19.81 (8.72)	5594.81	27.20
Overseas living purpose	SQ12R3	210.05 (25.77)	30.00 (4.49)	5594.81	27.20

From the estimated values of $\hat{\gamma}_{10}$ in the table, one can also see that each background variable was significantly associated with examinees' scores. The specific relationships of the four nominal background variables with TOEIC Listening scores are described here:

For the predictor education level, compared with the examinees who had vocational/technical school and community/junior college education levels (with average score of 234.15), all other examinees were more proficient by 56.50:

1. For the predictor current occupation, compared with the full-time employed examinees (with average score of 253.71), the part-time employed examinees were more proficient by 8.65; the full-time student examinees, by 34.60; and the other examinees (those unemployed and with missing response on this question), by 26.79.

2. For the predictor overseas living purpose, compared with examinees who did not respond to this question (with the average score of 267.76), the examinees living overseas for travel, work, and other purposes were more proficient by 24.63; those living overseas for study in an English program, by 64.03; and those living overseas for study in a non-English program, by 117.35.

3. For the predictor test-taking purpose, compared with examinees taking the test for promotion (with average score of 241.44), the examinees taking the test for English-language program evaluation were more proficient by 69.26; the examinees taking the test for course graduation, by 55.70; and the examinees with any other purpose, by 38.58.

4. For the four ordinal background variables, the examinees' scores were higher by 19.69, 18.00, and 37.47 when English study time, daily English use time, and overseas English experience increased by one level respectively. Examinees' scores were lower by 28.96 when self-reported English communication difficulty increased by one level.

It should be noted that the relationships of examinees' background variables with their TOEIC Listening scores were based on the average estimates across administrations. The standard deviations of slope estimates ($\hat{\gamma}_{10}$) in Table 8 show that the extent of the relationships varied across the 71 administrations. In other words, significant random effects were present in those relationships. On the other hand, although the extent of the relationships (i.e., the values of slopes) was different across administrations, the direction of the relationships (i.e., the signs of the slopes) was very consistent except for the background variable current occupation (i.e., the difference between full-time employed and part-time employed varied in both amount and direction across administrations, with 95% of the differences ranging from 8.65 - 6.13 * 1.96 = -3.36 to 8.65 + 6.13 * 1.96 = 20.66). The most stable and consistent relationship was between English study time and TOEIC Listening scores, with a slope of 19.69 and standard deviation of 1.10.

To evaluate the predictive power of each background variable on examinees' scores, the proportion of scores' variance explained by the background variable was calculated by

$$\frac{\hat{\sigma}^2(randomANOVA) - \hat{\sigma}^2(background)}{\hat{\sigma}^2(randomANOVA)}.$$

Based on the one-way ANOVA with random effect, $\hat{\sigma}^2$ = 7684.97. The $\hat{\sigma}^2$ column in Table 8 shows $\hat{\sigma}^2$ estimates based on different random-coefficient regression models. The last column in the table shows the percents of scores' variance explained by different background variables. For example, for the background variable based on education level, the explained variance percentage was (7684.97 – 7357.89) / 7684.97 = 4.26%. From the table, one can see that a single group composition variable could predict from 1.59% to 15.35% of the scores' variance. The best single predictor was overseas English experience.

The bottom row in Table 8 shows the results from the best combined-predictor model, which was identified based on variance explained and statistical significance of each predictor in the model. Based on this model, the best combined background variables for TOEIC Listening scores were English study time, English communication difficulty, overseas English experience, education level, current occupation, and overseas living purpose. Together, these six background variables explained 27% of scores' variance (i.e., *R* square = 0.27), and the RMSE was $\sqrt{5594.81}$ = 74.80. Therefore, examinees' background variables could not strongly predict their TOEIC Listening scores, although there were significant relationships between them.

Reading. Table 9 summarizes results from the eight single-predictor models and from the best combined-predictor model for TOEIC Reading scores. All the intercepts and slopes and their variances were statistically significant with $p < 0.001$ (not shown in the table). As the estimated values $\hat{\gamma}_{10}$ in the table show, each selected background variable had a strong relationship with examinees' scores. The specific associations of the four nominal background variables with TOEIC Reading scores are described are as follows:

1. For the predictor, education level, compared with the examinees who had vocational/technical school and community/junior college education levels (with an average score of 172.43), all other examinees were more proficient by 64.31.
2. For the predictor, current occupation, compared with the full-time employed examinees (with an average score of 209.26), the part-time employed examinees' were less proficient by 3.01; the full-time student examinees were more proficient by 22.94; and the other examinees (those unemployed and those with missing response on this question), by 16.01.
3. For the predictor, overseas living experience, compared with examinees with missing response on the question (with an average score of 216.09), the examinees living overseas for travel, work, and other purposes were more proficient by 18.98; those living overseas for study in an English program, by 45.88; and those living overseas for study in a non-English program, by 105.79.
4. For the predictor, test taking purpose, compared with examinees taking the test for promotion (with an average score of 197.92), the examinees taking the test for English-language program evaluation were more proficient by 50.01; those taking the test for course graduation, by 48.20; and those with any other purposes, by 26.09.

Table 9

Results From Random-Coefficient Regression Models for the TOEIC Reading Section

$$Y_{ij} = \beta_{0j} + \beta_{1j}B_{1ij} + r_{ij}\ ,\ \beta_{0j} = \gamma_{00} + u_{0j}\ ,\ \beta_{1j} = \gamma_{10} + u_{1j}$$

Level 1 predictor	Variable name	$\hat{\gamma}_{00}$ (SD)	$\hat{\gamma}_{10}$ (SD)	$\hat{\sigma}^2$	Variance explained %
Education level	SQ1R1	172.43 (10.65)	64.31 (7.63)	8830.99	4.61
Current occupation	SQ3R1	209.26 (8.01)	-3.01 (6.63)	9172.24	0.92
	SQ3R2		16.01 (8.22)		
	SQ3R3		22.94 (14.58)		
English study time	SQ6	162.24 (12.51)	22.54 (1.91)	8836.30	4.55
Daily English use time	SQ8	204.04 (16.15)	16.37 (3.65)	9093.89	1.77
English communication difficulty	SQ10	329.05 (31.25)	-26.28 (5.84)	8596.40	7.14
Overseas English experience	SQ11	184.73 (18.20)	30.64 (3.40)	8468.66	8.52
Overseas living purpose	SQ12R1	216.09 (15.80)	18.98 (3.26)	8641.54	6.65
	SQ12R2		45.88 (7.31)		
	SQ12R3		105.79 (12.50)		
Test-taking purpose	SQ14R1	197.92 (11.86)	26.09 (11.85)	9109.29	1.60
	SQ14R2		48.20 (19.38)		
	SQ14R3		50.01 (13.88)		
English study time	SQ6	150.15 (21.12)	19.89 (1.62)	7409.45	19.96
English communication difficulty	SQ10	150.15 (21.12)	-16.74 (4.15)	7409.45	19.96
Overseas English experience	SQ11	150.15 (21.12)	19.53 (3.27)	7409.45	19.96
Education level	SQ1R1	150.15 (21.12)	50.26 (5.19)	7409.45	19.96
Current occupation	SQ3R3	150.15 (21.12)	11.92 (10.62)	7409.45	19.96
Overseas living purpose	SQ12R3	150.15 (21.12)	38.13 (4.90)	7409.45	19.96

For the four ordinal background variables, the examinees' TOEIC Reading scores were higher by 22.54, 16.37, and 30.64 when English study time, daily English use time, and overseas English experience increased by one level respectively. Examinees' scores were lower by 26.28 when the self-reported English communication difficulty increased by one level.

As mentioned in the TOEIC Listening section, the relationships of examinees' background variables with their TOEIC Reading scores were based on the average estimates across administrations. The standard deviations of slope estimates ($\hat{\gamma}_{10}$) in Table 9 show the significant random effects in those relationships and the extent of the relationships varied across the 71 administrations. Again, although the extent of the associations (i.e., the values of slopes) was different, the direction of the associations (i.e., the signs of the slopes) was very consistent across administrations except for the background variable current occupation (i.e., the differences between subgroups varied in both amount and direction across administrations, with 95% of differences in the ranges of -3.01 ± 1.96 * 6.63, 16.01 ± 1.96 * 8.22, and 22.94 ± 1.96 * 14.58, respectively). The most stable and consistent

association was again the relationship of English study time with TOEIC Reading scores, with a slope of 22.54 and standard deviation of 1.91.

Based on the last column in Table 9, one single group composition variable could predict from 0.92% to 8.52% of the TOEIC Reading scores' variance. The best single predictor was overseas English experience. The bottom row in Table 9 shows the results from the best combined-predictor model identified in the analysis. The best combined background variables for TOEIC Reading scores were English study time, English communication difficulty, overseas English experience, education level, current occupation, and overseas living purpose. These six background variables together explained 20% of the scores' variance (i.e., *R* square = 0.20), and the RMSE was $\sqrt{7409.45}$ = 86.08. Therefore, examinees' background variables could not strongly predict their TOEIC Reading scores, although there were significant relationships between them.

Therefore, the results from random-coefficient regression models suggest the presence of close relationships between examinees' background and their test scores for both sections. Using six background variables as predictors, around 27% and 20% of scores' variance could be explained for the TOEIC Listening and Reading sections. In other words, the correlation coefficient between test scores and examinees' background was 0.52 and 0.45 for the two sections. The prediction error was 75 for the TOEIC Listening section and 86 for the TOEIC Reading section.

When the interactions between the six background variables were added in the best models, the *R* square increased by 0.003 for the TOEIC Listening section and by 0.0038 for the TOEIC Reading section and the RMSE decreased by 0.12 for the TOEIC Listening section and by 0.21 for the TOEIC Reading section. Therefore, adding interactions in the models did not improve the prediction.

A regular regression model with examinees' test scores as the dependent variable and their background as the independent variables (i.e., using only Level 1 data) was also conducted, and the same model previously identified was still the best. For both the TOEIC Listening and Reading sections, the regular regression model was similar to the multilevel regression with random-effects model in terms of predictors, regression coefficient estimates, *R* square and adjusted *R* square, and RMSE, with the latter's prediction error being slightly lower. However, the multilevel analysis was more functional because it showed the variability of relationships between examinees' background and their scores across administrations.

Using Background Information to Predict Test Performance: Validation

Based on the best random-coefficient regression models identified in the study, at the individual examinee level, about 27% and 20% of scores' variance could be explained by examinees' background variables, and the RMSEs were 75 and 86 for the TOEIC Listening and Reading sections, respectively. Given that the examinees' scores ranged from 10 to 495 and from 5 to 495, with standard deviations of 89 and 97 for the two sections (see Table 1), the prediction was not strong. Based on the best regression with means-as-outcomes models at the administration level for the TOEIC Listening and Reading sections, about 85% and 84% of the means' variance could be

explained by group composition variables, and the RMSE was about 6. Given that the scale score means varied from 256 to 322 for the TOEIC Listening section and from 202 to 258 for the TOEIC Reading section with a standard deviation of 16 (see Table 1), the prediction was very strong.

Therefore, the regression with means-as-outcomes models has the potential to predict scale score means from examinees' group composition for future administrations. To confirm and validate the prediction models, the group composition data from 14 new administrations were collected and the predicted means were calculated. The predicted means were then compared with the observed means based on operational equating and scoring. Table 10 shows the predicted means, operational means, and their differences for both the TOEIC Listening and Reading sections. Compared with operational means, the predicted means were higher on some administrations but lower on others. The mean differences were from -9.83 to 11.13 for the TOEIC Listening section and from -5.99 to 21.42 for the TOEIC Reading section. For both sections, 71% (i.e., 10/14) of the absolute values of mean differences were smaller than the RMSE (i.e., 6) estimated in the prediction models. This finding is consistent with the finding from the models that 68% of the actual means should be within ±6 range of the predicted means for both the TOEIC Listening and Reading sections. Although the differences varied across administrations and test sections, the average differences were very small (i.e., 0.79 for the TOEIC Listening section and 4.04 for the TOEIC Reading section). Therefore, the prediction models for scale score means were confirmed and validated.

Table 10

New Administrations' Predicted Means and Operational Means

New admin	Listening mean			Reading mean		
	Predicted	Operational	Predicted-operational	Predicted	Operational	Predicted-operational
1	311.11	311.20	-0.09	250.25	252.60	-2.35
2	316.16	312.60	3.56	256.85	252.70	4.15
3	303.83	298.70	5.13	246.99	234.70	12.29
4	293.94	292.50	1.44	235.45	232.80	2.65
5	278.61	278.40	0.21	219.43	202.40	17.03
6	286.03	274.90	11.13	224.34	218.70	5.64
7	297.17	296.20	0.97	241.54	240.70	0.84
8	316.53	317.20	-0.67	257.96	262.40	-4.44
9	316.94	315.11	1.83	262.80	241.38	21.42
10	291.64	282.98	8.66	232.18	227.87	4.31
11	291.19	292.30	-1.11	241.81	236.43	5.38
12	277.33	284.90	-7.57	220.77	225.09	-4.32
13	278.78	281.35	-2.57	221.59	221.59	0.00
14	291.05	300.88	-9.83	232.31	238.30	-5.99
Mean	296.45	295.66	0.79	238.88	234.83	4.04
SD	14.34	14.25	5.53	14.64	15.50	8.11
Minimum	277.33	274.90	-9.83	219.43	202.40	-5.99
Maximum	316.94	317.20	11.13	262.80	262.40	21.42

Discussion and Conclusions

Different methods and techniques have been proposed to monitor scaled scores across test administrations (von Davier, 2012), and the association of examinees' background with their test performance has also been explored in some studies (Liu et al., 2012; Luo et al., 2011; Wei & Qu, 2012). This study used multilevel analysis models to explore the relationships between examinees' background and their scaled scores on the TOEIC Listening and Reading test at both the individual and administration levels. The models with strong predictive power were then applied to later operational data to confirm their validity.

The study started with carefully defining and coding examinees' responses to a background questionnaire based not only on the specific questions and their options but also on the performance consistency of subgroup examinees that chose each option. This first step was important for further analyses. For some unknown reason (e.g., misunderstanding, cultural difference, sensitivity, or motivation), examinees may respond in unexpected ways to some background questions; this can lead to unexpected yet consistent performance for some background question options. From this study, even those examinees with missing data on some background questions had consistently lower or higher performance across administrations. In addition, for some questions, examinees who chose different options performed similarly on all or most test administrations. This coding rationale is especially important in creating Level 1 categorical background variables in this study. For group composition variables at Level 2, those examinees choosing different but close options might be combined together to obtain a new subgroup's percentage because the new subgroup's percentage has a stronger relationship with test score means than separate subgroups' percentages. The purpose of this coding strategy is to effectively and fully use examinees' background information for statistical analysis without loss of parsimony. Compared with this coding strategy, the study by Liu et al. (2012) used examinees' original responses to background questions in their statistical analysis. The study by Luo et al. (2011) used imputation and deletion to handle missing background information and then defined the sample sizes of cross-classification groups as the background composition variables in their analysis.

Based on the one-way ANOVA model with random effects, the grand mean estimates for all examinees were very precise, and the group mean estimates were very reliable. The low score dependence within administrations suggests that examinees should not be concerned about which administrations or forms to take. However, score means fluctuated substantially across administrations. From a quality control perspective, this might signal either population changes across administrations or equating problems due to large differences in test form difficulty, large differences between groups, or differential performance of anchor sets between the two groups. The multilevel analyses first focused on exploring the relationship between means fluctuation and population change across administrations.

The results from the regression with means-as-outcomes model suggested that each of 10 group composition variables had a strong relationship with group means and could separately account for 20–60% of the means' variance. For both the TOEIC Listening and Reading sections, five group

composition variables could together account for 84–85% of means variance, with the prediction error of 6 scale score points. Therefore, the group composition variables were very powerful predictors of administration means. The prediction models were validated by using the data from 14 new administrations. The results provide strong empirical support for the hypothesis proposed but not proved in the study by Luo et al. (2011; i.e., "the variation in the examinee composition across administrations is a major reason for fluctuations in the mean of the scaled scores" p. 2). In addition, this study found that the seasonality of scale score means could be fully accounted for by the examinee composition variables. However, it should be noted that the examinee composition variable in this study was defined as the percentages of some carefully selected subgroup(s), instead of the sample sizes of cross-classification groups, which was used in the study by Luo et al. In terms of the prediction accuracy, the prediction errors for scale score means from this study were about half as large as those from the study by Liu et al. (2012), which used similar test data and background information. One of the main reasons the two previous studies could not find a strong prediction model was the small numbers of administrations (i.e., 10 and 15 administrations) that were examined.

Compared with the very positive results at Level 2, the prediction of individuals' scores based on their background variables was not strong, although most background variables separately or collectively had significant relations to test scores. Based on the random-coefficient regression model, six variables together could explain 27% of TOEIC Listening scores' variance and 20% of the TOEIC Reading scores' variance, with the prediction error of 75. The significant random effects of both intercepts and slopes suggest that the subgroups' performance and their difference changes with administrations. In operational work, it is not unusual to use subgroups' performance in previous administrations and subgroups' sample sizes in the current administration to predict current test performance. The finding from this study indicates that, at least for this test, it may not be appropriate to weight subgroups' average scores by their frequencies to predict or verify an administration's test performance.

The finding of a stronger prediction model at the administration level and a weaker prediction at the examinee level is not surprising given the different types of variables and units of analysis at the two levels. The Level 1 predictors were basically categorical variables, and the Level 2 predictors were ordinal variables (i.e., percentages). The unit of analysis at Level 1 was an individual examinee's information (i.e., test scores and background), and the unit of analysis at Level 2 was an administration's accumulative information (i.e., test score means and group composition). This pattern of lower prediction at the examinee level and higher prediction at the administration level is similar to the finding from the multilevel analysis of an English speaking test scores (Wei & Qu, 2012) that 34% of the administration means' variance could be predicted by two group composition variables and that 21% of the individual scores' variance could be explained by four background variables. The predictive power from the current study is much stronger than that from the previous study especially at the administration level. The following factors may be related to the difference in predictive power: (a) The tests in the two studies measured different types of constructs or abilities, with one examining productive language skill (i.e., speaking) and the other examining receptive

skills (i.e., the TOEIC Listening and Reading test); (b) the tests were different in item format (i.e., multiple-choice items versus constructed-response items), scoring (i.e., objective and automatic scoring versus subjective and human scoring), and length (i.e., 100 items versus 13 items); (c) the comparability of scores across administrations/forms were controlled by different ways (e.g., equating versus human scoring); and (d) the examinees in this study were first-time test takers, while the examinees in the other study included repeaters, which may have led to dependence of scores across administrations.

In the operational work, the usefulness of examinees' background information depends on its relationship with examinees' test performance. The stronger the relationship, the more useful it will be. For the *TOEIC*® program used in this study, for both the TOEIC Listening and Reading sections, the relationship between administration means and group composition is very close ($r = 0.92$); the predictive power is very strong (R square = 0.84 – 0.85), the prediction error is very small (RMSE = 6), and the prediction model is effectively validated. Therefore, the examinees' background information will be very useful for the operational work. The predicted scale score means based on the background information can not only be used to predict and understand test performance before and after test scoring, they can also be informative and useful in making the selection of equating method. It is well known that score equating is essential for almost any testing program with frequent administrations, and some principles and procedures need to be followed (Dorans, Moses, & Eignor, 2010). On the other hand, various challenges exist in real world equating. A testing program often tries different equating methods and then evaluates which method produces the most reasonable and consistent results. However, it is sometimes hard to say which method produces more accurate scale score means on a specific administration because it is hard to empirically check the underlying assumptions of the different equating methods in operational work. Sometimes the fluctuation of scale score means across administrations makes the judgment even more challenging. When it is unclear which equating method provides the most accurate results, some external information outside of equating (e.g., examinees' background or group composition) may provide useful information about the group's performance level. In other words, the predicted means based on examinees' group composition may provide external clues or validity for selecting one of the equating methods for scoring. In situations when all equating methods produce unusually high or low administration means, the examinees' background information and its predicted means can help understand the group's performance level. Therefore, it should be a part of operational quality control procedures to identify the relationship between examinees' background information and their test scores, and then use the relationship to understand and monitor test performance across administrations.

The results from this study have practical implications for the testing program. As a flexible method, the multilevel analysis can be used to identify statistical relationships between examinees' background and their test scores at both the individual examinee level and the administration level. In operational work, examinees' demographic information can be used to evaluate the consistency and variability of subgroups' performance across different administrations. Group composition can be used to predict and monitor changes in examinees' performance over administrations. For future administration, the prediction error from the model can be used to construct a confidence interval

for the predicted mean, which can then be used to evaluate whether the observed mean falls within that interval. The finding of the strong relationships between examinees' background and their test scores also provide empirical evidence for the validity of the TOEIC Listening and Reading test. Therefore, the procedures and results from this study are an important part of the quality assurance process for the testing program.

In sum, based on the multilevel analysis of the data collected from 71 administrations of the TOEIC Listening and Reading test, this study found the following for both sections: (a) at the examinee level, examinees' background information had strong relations to their test scores and the relations varied across administrations; however, the prediction of individuals' test scores based on their background variables was not strong; and (b) at the administration level, group composition had strong relations to administration means; the prediction of administration means based on group composition variables was very strong and the model had potential application in understanding and monitoring the TOEIC test performance across administrations. The findings from this study also indicate that it may be very helpful for a testing program to monitor test performance across administrations if a well-designed questionnaire is used to collect examinees' background information during the registration or administration of the test.

Limitations and Future Research

The results from this study are very positive and promising. The strong prediction models for scale score means based on examinees' background information have significant potential in the quality control of the TOEIC Listening and Reading test. However, there are some limitations in this study and future research may address these limitations.

First, the data used in the models may not be ideal to fully explore true relationships. For the dependent variables, the test scores were rounded and truncated scale scores with an interval of 5; for the independent variables, the background information was represented by categorical and ordinal variables. This may have attenuated their relationships in statistical modeling and estimation. In addition, the relatively small number of Level 2 units in the analyses may have an impact on the accuracy of random effect estimation for administrations.

Second, the prediction of the two sections' scores in the test, the TOEIC Listening and Reading sections, was explored separately, and the relationship between section scores was ignored in the analyses. In the future, multivariate multilevel analysis could be used to simultaneously explore relationships with examinees' background, with the consideration of the two sections' relationship in the model.

Third, it is well known that while using a regression model to predict a criterion variable, the phenomenon of *regression toward the mean* effect is present as long as a less than perfect correlation exists between the criterion variable and predictor(s). Although the regression effect is not serious at the administration level in this study due to the high correlations between scale score means and predictors, one needs to be careful while interpreting the predicted scale score means especially for those administrations with very high or low performance levels.

Fourth, the relationship between examinees' background and their test performance may vary depending on the type of background information, the format of the test, the underlying construct the test is designed to measure, the population taking the test, and the quality of equating and scoring. More studies need to be conducted to explore the relationship in different testing programs.

Finally, for a testing program with frequent administrations, the relationship of examinees' background with their test performance may gradually change. Therefore, it is necessary to reexamine the relationship and adjust the prediction model during the life of a testing program.

References

Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice, 26*, 36–46.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (Research Report No. RR-10-29). Princeton, NJ: Educational Testing Service.

Haberman, S., & Dorans, N. J. (2011). *Sources of score scale inconsistency* (Research Report No. RR-11-10). Princeton, NJ: Educational Testing Service.

Haberman, S., Guo, H., Liu, J., & Dorans, N. J. (2008). *Consistency of SAT® I: Reasoning test score consistency* (Research Report No. RR-08-67). Princeton, NJ: Educational Testing Service.

Kasim, R., & Raudenbush, S. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous with-in group variance. *Journal of Educational and Behavioral Statistics, 20*, 93–116.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education, 3*, 97–104.

Lee, Y.-H., & Haberman, S. (2011). *Application of harmonic regression to monitor scale stability*. Manuscript in preparation.

Lee, Y.-H., & von Davier, A. A. (in press). Monitoring scale scores over time via quality control tools and time series techniques. *Psychometrika*.

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-Verlag.

Liao, C. W., & Livingston, S. A. (2012, April). *A search for alternatives to common-item equating*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Liu, M., Lee, Y.-H, & von Davier, A. A. (2012, July). *Detection of unusual administrations using a linear mixed effects model*. Paper presented at the international meeting of the Psychometric Society, Lincoln, NE.

Luo, L., Lee, Y.-H., & von Davier, A. A. (2011, April). *Pattern detection for scaled score means of subgroups across multiple test administrations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Paek, I., Liu, J., & Oh, H. J. (2010). *An investigation of propensity score matching on linear/nonlinear observed score equating*. Unpublished manuscript.

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

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2000). HLM 6 *hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

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.

Visser, I., Raijmakers, M. E. J., & van der Maas, H. L. J. (2009). Hidden Markov models for individual time series. In J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, & N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 269–289). New York, NY: Springer.

Wei, Y., & Qu, Y. (2012, April). *Using multilevel analysis to monitor test performance across administrations*. Paper presented at the annual meeting of National Council on Measurement in Education (NCME), Vancouver, British Columbia, Canada.