Listening. Learning. Leading.®

Research Report
ETS RR–14-29

# Using Multilevel Analysis to Monitor Test Performance Across Administrations

Youhua Wei

Yanxuan Qu

December 2014

# 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

# Using Multilevel Analysis to Monitor Test Performance Across Administrations

Youhua Wei & Yanxuan Qu

Educational Testing Service, Princeton, NJ

For a testing program with frequent administrations, it is important to understand and monitor the stability and fluctuation of test performance across administrations. Different methods have been proposed for this purpose. This study explored the potential of using multilevel analysis to understand and monitor examinees' test performance across administrations based on their background information. Based on the data of 330,091 examinees' test scores and their background information collected from 254 administrations of an English-speaking test, the study found: (a) at the individual examinee level, examinees' background had statistically significant relationships with their test performance, and the relationships 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 relationships with administration means; the prediction of administration means based on group composition variables was fairly strong. The results suggest that multilevel analysis has potential application in understanding and monitoring test performance across administrations by identifying statistical relationships between examinees' characteristics and their test performance at both individual and administration levels.

**Keywords** Multilevel analysis; background information; test performance; quality control; prediction model

doi:10.1002/ets2.12029

For a testing program with many forms and administrations, test performance may fluctuate over time, even though efforts have been 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. 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 (e.g., Dorans, 2004; Haberman, Guo, Liu, & Dorans, 2008). 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 the following information: examinees' background change, subpopulation shift, seasonality of the test performance, scale shift, test difficulty shift, and so on. More and more 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, 2013), multivariate mixed weighted modeling (Luo, Lee, & von Davier, 2011), linear mixed effects modeling (Lee, Liu, & von Davier, 2013), Shewhart control charts (see a brief description in von Davier, 2012), and hidden Markov model (Lee & von Davier, 2013).

Multilevel analysis (i.e., hierarchical linear modeling, Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) is useful for understanding relationships among variables that exist at different levels in a hierarchical data structure. With its strength in modeling both fixed and random effects, this method has been widely used in behavioral, educational, and other social sciences. In the past decades, multilevel analysis has been used in educational measurement with the combination of item response theory (IRT) to understand examinees' performance on individual items (Adams, Wilson, & Wang, 1997; Boeck & Wilson, 2004; Rijmen, Tuerlinckx, Boeck, & Kuppens, 2003). However, the multilevel or hierarchical data structure is typically defined in a traditionally *natural* way (e.g., examinees nested in classes, schools, gender, or social economical statuses). In a testing program with many administrations, the test data can be considered to have a two-level hierarchical structure, with the individual examinees at Level 1 and the administrations at Level 2. In this data structure, the examinees'

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

characteristics (e.g., education level, occupation, gender, and social economic status) are considered as Level 1 variables, and test administration information (years or months, test forms, population characteristics) are considered as Level 2 variables. Then the individual examinees' test scores can be predicted by their demographic information; the test form or administration score means can be understood from administration information; and the stability and fluctuation of test performance across administrations can be monitored by relationships among examinee-level and administration-level variables.

It is not unusual for a testing program to collect examinees' background information during the registration or administration of the test. The background information may include general demographic information such as race, gender, education, vocation, and socioeconomic status. It may also include examinees' training, learning, or test-taking experiences, which are related to the construct the test is designed to measure. Both empirical research (e.g., Lee et al., 2013; Luo et al., 2011; Wei, 2013) and operational experience suggest that there are relationships between examinees' test scores and their background. Some studies have been conducted to explore the potential of using examinees' background information to improve equating accuracy. Although a recent study found that it was useful to adjust group composition for IRT linking and equating (Qian, von Davier, & Jiang, 2013), some other studies concluded that examinees' background information did not provide additional information for equating (e.g., Kolen, 1990; Paek, Liu, & Oh, 2010). It was not recommended to adjust group ability difference for equating purposes (Liao & Livingston, 2012). However, examinees' background information still has the potential in monitoring test performance. As Allalouf (2007) suggested, it should be part of the quality control procedure for a testing program to explore the statistical relationship between examinees' background and their scores and then use the relationship to understand and monitor test scores. Two studies (Lee et al., 2013; Luo et al., 2011) have been conducted for this purpose. However, the small numbers of administrations used in those studies (i.e., 10 and 15 administrations) 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 would impact examinees' scores. More importantly, how well the prediction models could be used to understand and monitor test performance for future administrations was not fully examined. A recent study (Lee & Haberman, 2013) found that changes in the regional distribution of examinees (defined by test center countries) well explained the variability in the mean scores of an international language assessment.

This study investigated the potential of using multilevel analysis to understand and monitor test performance over administrations based on examinees' background information. Specifically, the study used a large-scale data set with a big number of administrations to address the following questions:

- How strong are the relationships between examinees' test scores and their background information at the individual level?
- Are those relationships consistent or varied across different administrations?
- How strong are the relationships between examinees' mean scores and their background information at the administration level?
- Can the relationships be used to understand and monitor test performance across administrations?

## Methodology

### Data

The data for this study were collected from an English-language testing program in a country where English is the second language. The test was designed to evaluate examinees' English-speaking skills by 13 constructed response items. A weighting method was used to compute the total test raw scores (ranging from 0 to 24), based on the importance of different tasks to the speaking skills. Given that the comparability of test scores on different forms was controlled and monitored by using consistent form development procedures and the same scoring rubrics, the raw scores were used as examinees' test scores in this study.

A background questionnaire was used to collect examinees' information on education and work-related background, English-language experience, and test-taking experience. Specifically, there were 14 questions in the questionnaire with different options: five questions about examinees' education and 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 about

examinees' English-language experience (e.g., How many years have you spent studying English?), and two questions about examinees' experience in taking the test (e.g., What is your main purpose for taking today's test?).

The data used in this study include 330,091 examinees' scores and their background information, collected from 254 administrations of the test in 4 years. The examinees with missing information on test scores or any background questions were not included in the data. With each test form being used in each administration, the sample sizes range from 260 to 12,389, with an average of 1,300. The administration score means range from 13.29 to 16.76, with an average of 14.99. The individual examinees' scores across administrations range from 0 to 24, with an average of 15.04 and standard deviation of 2.86.

## Procedure and Analysis

### *Data Preparation*

The data of all examinees' test scores and their responses to the 14 background questions were reorganized at two levels. At Level 1, the test scores of individual examinees were used as the dependent variable. The coding of examinees' background information as predictor variables was based on each background question's original response options. If the frequencies of the responses to some options were very small and the options were adjacent or close to each other, a new code would be created to represent the combined options. For example, for the background of *education level*, the examinees with education levels below undergraduate were combined together as a new subgroup in the coding (see Table 1). At Level 2, the test score means of each administration were used as the dependent variable. The predictors were group composition variables, which were defined as the percentages of subgroups based on examinees' responses to background questions in specific administrations (see Table 2). The following section describes in more detail how the background variables were coded and selected at both levels.

### *Preliminary Analyses and Variable Selection*

To explore the relationships between examinees' test performance and their background information, it was important to select important background questions carefully 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 score means of subgroups based on response options were plotted and compared across administrations. If there was a consistent pattern of score means between subgroups across administrations, the background question was selected for further analyses. For example, for the background *test-taking purpose*, there was a consistent test performance pattern among the five subgroups (i.e., the examinees taking the test for *job promotion* tended to have lower score means than other examinees), so this background question was chosen for further analyses. At Level 2, the bivariate scatter plots and correlations between group composition variables and test score means were used to explore their relationships. The decision of which subgroups would be used to define the group composition variables for further analyses was based on their bivariate relationships with test score means. If the percentage of combined subgroups had stronger correlation with test score means than the percentage of any single subgroup, the group composition variable would be based on the combined subgroups. For example, for the background question, *How often has difficulty with English affected your ability to communicate?*, the group composition variable based on the combined responses of *never*, *seldom*, and *sometimes* had stronger association with test score means than group composition variables based on any single subgroups, so it was used as the final composition variable for this background question.

Based on the preliminary analyses, eight background questions were selected to create independent variables at both Level 1 and Level 2 for this study (see Tables 1 and 2 for detailed information about the selected background questions, coded variables, and their summary statistics). At Level 1, four questions were coded as categorical variables and four as ordinal variables, based on the scaling feature of those questions. Each of the four questions with categorical variables was dummy coded with one subgroup as the reference group. The four ordinal variables were considered as having interval scales in the multilevel analyses, although the intervals between adjacent values of the four ordinal variables may be not equal (e.g., *almost never*, *seldom*, *sometimes*, *frequently*, and *almost always*). At Level 2, the percentages of selected subgroups based on those eight questions were used as group composition variables at the administration level.

**Table 1** Level 1 Variables and Codes ($N = 330,091$)

| Background question | Option | Code | Subgroup percent | Variable |
|---|---|---|---|---|
| Education level | College and below | (0, 0, 0) | 5.15 | Reference |
| | Undergraduate | (1, 0, 0) | 84.22 | Edu1 |
| | Graduate | (0, 1, 0) | 9.78 | Edu2 |
| | Language institute | (0, 0, 1) | 0.91 | Edu3 |
| Occupation | Full-time employed | (0, 0, 0) | 20.22 | Reference |
| | Part-time employed | (1, 0, 0) | 4.78 | Occ1 |
| | Unemployed | (0, 1, 0) | 13.69 | Occ2 |
| | Full-time student | (0, 0, 1) | 61.31 | Occ3 |
| English study time | ≤4 years | 1 | 4.36 | Stu |
| | 4 – 6 years | 2 | 6.76 | |
| | 6 – 10 years | 3 | 29.75 | |
| | >10 years | 4 | 59.13 | |
| English use time | None | 1 | 5.51 | Use |
| | 1 – 10% | 2 | 43.02 | |
| | 11 – 20% | 3 | 29.81 | |
| | 21 – 50% | 4 | 16.11 | |
| | 51 – 100% | 5 | 5.54 | |
| English difficulty | Almost never | 1 | 2.61 | Dif |
| | Seldom | 2 | 10.61 | |
| | Sometimes | 3 | 47.42 | |
| | Frequently | 4 | 29.08 | |
| | Almost always | 5 | 10.27 | |
| Overseas English experience | None | 1 | 41.21 | Ove |
| | <6 months | 2 | 26.49 | |
| | 6 – 12 months | 3 | 21.00 | |
| | 1 – 2 years | 4 | 6.60 | |
| | >2 years | 5 | 4.70 | |
| Test-taking time | Never | (0, 0, 0) | 55.57 | Reference |
| | Once | (1, 0, 0) | 24.22 | Tim1 |
| | Twice | (0, 1, 0) | 9.36 | Tim2 |
| | Three or more | (0, 0, 1) | 10.85 | Tim3 |
| Test-taking purpose | Promotion | (0, 0, 0, 0) | 10.70 | Reference |
| | Job application | (1, 0, 0, 0) | 68.48 | Pur1 |
| | Program evaluation | (0, 1, 0, 0) | 4.48 | Pur2 |
| | Learning evaluation | (0, 0, 1, 0) | 11.97 | Pur3 |
| | Course graduation | (0, 0, 0, 1) | 4.37 | Pur4 |
| Test score | Mean = 15.04, $SD = 2.86$, Minimum = 0, Maximum = 24 | | | |

*Note*. Edu = education; Occ = occupation; Stu = study; Dif = difficulty; Ove = overseas; Tim = time; Pur = purpose.

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 relations of examinees' background to their test performance across administrations, with examinees at Level 1 and administrations at Level 2. Based on the preliminary analyses, different models were explored, and results were evaluated in terms of the prediction of test performance based on examinee's background at both levels. Specifically, four models were used in the study:

The first model is one-way ANOVA model with random effect

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

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

**Table 2** Level 2 Variables and Summary Statistics ($N = 254$)

| Background | Group composition (Percent of subgroup) | Variable | Mean | *SD* | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Education level | Undergraduate | Gedu | 83.85 | 3.94 | 61.6 | 91.3 |
| Occupation | Full-time employed | Gocc | 21.88 | 11.36 | 6.7 | 60.5 |
| English study time | 6–10 years | Gstu | 29.55 | 2.73 | 21.6 | 44.3 |
| English use time | <20% | Guse | 77.97 | 3.44 | 59.1 | 84 |
| English difficulty | Never, seldom, sometimes | Gdif | 60.72 | 3.62 | 53.6 | 78.3 |
| Overseas English experience | 0–6 months | Gove | 67.72 | 5.37 | 44.9 | 79.3 |
| Test-taking time | Never | Gtim | 54.13 | 8.77 | 31 | 85.9 |
| Test-taking purpose | Promotion | Gpur | 11.77 | 8.19 | 1.8 | 46 |
| Score means | | Sample | 1,299.57 | 858.45 | 260 | 12, 389 |
| | | Mean | 14.99 | 0.57 | 13.29 | 16.76 |
| | | *SD* | 2.8 | 0.27 | 2.15 | 3.61 |

*Note*. Gedu = group education; Gocc = group occupation; Gstu = group study; Guse = group use; Gdif = group difficulty; Gove = group overseas; Gtim = group time; Gpur = group purpose.

where $Y_{ij}$ is the test score of examinee $i$ on administration $j$; $\beta_{0j}$ 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 assumed to be normally distributed with $N(0, \sigma^2)$; $\gamma_{00}$ 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 assumed to be normally distributed with $N(0, \tau_{00})$.

The second model is regression with means-as-outcomes

$$\text{Level 1}: Y_{ij} = \beta_{0j} + r_{ij};$$
$$\text{Level 2}: \beta_{0j} = \gamma_{00} + \gamma_{01} G_j + u_{0j},$$

where $G_j$ is the Level 2 predictor or group composition variable for score mean on administration $j$; $\gamma_{01}$ is the slope in regression of $\beta_{0j}$ on predictor $G_j$; $\gamma_{00}$ 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 $N(0, \tau_{00})$.

The third model is random-coefficient model

$$\text{Level 1}: Y_{ij} = \beta_{0j} + \beta_{1j} B_{1ij} + r_{ij};$$
$$\text{Level 2}: \beta_{0j} = \gamma_{00} + u_{0j}, \beta_{1j} = \gamma_{10} + u_{1j},$$

where $B_{1ij}$ is the predictor or examinee's background variable at Level 1; $\beta_{0j}$, $\beta_{1j}$ are intercept and slope in regression of $Y_{ij}$ on Level 1 predictor; $r_{ij}$ is the residual conditioned on Level 1 predictor; $\gamma_{00}$ and $\gamma_{10}$ 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 $\tau_{00}$ is the unconditional variance in the Level 1 intercepts, $\tau_{11}$ is the unconditional variance in the Level 1 slopes, and $\tau_{01}$ or $\tau_{10}$ is the unconditional covariance between the Level 1 intercepts and slopes.

The fourth model is the intercepts- and slopes-as-outcomes model

$$\text{Level 1}: Y_{ij} = \beta_{0j} + \beta_{1j} B_{1ij} + r_{ij};$$
$$\text{Level 2}: \beta_{0j} = \gamma_{00} + \gamma_{01} G_j + u_{0j}, \beta_{1j} = \gamma_{10} + \gamma_{11} G_j + u_{1j},$$

where $G_j$ is the Level 2 predictor for the intercept and slope; $\gamma_{00}$ and $\gamma_{01}$ are the grand mean and slope for the Level 1 intercept; $\gamma_{10}$ and $\gamma_{11}$ are the average mean and slope for the Level 1 slope.

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 score means and group composition variables at Level 2. The random-coefficient model was used to explore the relationship between examinees'

test scores and their background at Level 1. Finally, we used intercepts- and slopes-as-outcomes model to check the consistency of the Level 1 background-score relationship across administrations.

For both the regression with means-as-outcomes model and random-coefficient model, the analyses started with one level, with the other level being held aside. Each background variable was first used as the single predictor in the model, so that the predictive power of each background for test performance could be evaluated. Then multiple background variables were included at the same level in the model to explore and identify the best prediction model for test performance. 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 the predictive power and accuracy. Each model was evaluated by the prediction coefficient(s) and the proportion of test score variance explained by the predictor(s). The prediction coefficient(s) was used to examine the specific relationship between the background and test performance, and the proportion of score variance explained by the predictor(s) was used to evaluate the predictive power of the background variable(s). The results from the best models identified in the preliminary analyses were also used to select the best predictors.

For the metrics 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 (see Table 1); at Level 2, the group composition variables were centered around their corresponding grand means (i.e., $G_j - \overline{G}$), based on the suggestion by Raudenbush and Bryk (2002). The setup of these locations 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 model with the strongest predictive power, examinees' background data from 22 other new administrations not included in the modeling were used to predict their test performance (e.g., group mean scores on those administrations). These predicted scores were then compared with test scores produced from operational scoring. The results were used to validate the prediction model.

## Results

In this section, we first summarize the one-way ANOVA results, which can provide baseline information for further analyses. Then we explore the prediction model for score means at the administration level by using regression with the means-as-outcomes model and the prediction model for test scores at the examinee level by using the random-coefficient regression model. We also use the intercepts- and slopes-as-outcomes model to examine how the score-background relationship changed across administrations. The section closes by applying the strongest prediction model identified in the analyses to the new operational data and evaluating its validity. To precisely show and compare the results from different models, we keep three decimals for most numbers in this section.

### Estimating Variance Components: One-Way ANOVA With Random Effect

As the simplest model in multilevel analysis, the one-way ANOVA model provides preliminary results about variation of the test scores within and between administrations. It also provides reliability of observed administration means (i.e., sample means) for the true means of the populations on those administrations.

Based on the results from the one-way ANOVA mode with random effect and homogeneity assumption of Level 1 variance $\sigma^2$ (see Table 3 for the detailed results), the grand mean of test scores across administrations was $\widehat{\gamma}_{00} = 14.995$, with a standard error of 0.036. So the 95% confidence interval for the grand mean of test scores was $14.995 \pm 1.96 * 0.036 = (14.924, 15.066)$. Although the reliability of sample mean as an estimate of the true mean on the administration may vary across administrations due to different sample sizes, an overall reliability estimate of the observed sample means $\widehat{\beta}_{0j}$ was $\widehat{\lambda} = 0.977$. Therefore, the grand mean estimate appeared to be very precise, and the sample means from administrations tended to be very reliable estimates for the true score means.

For the variance components, Level 1 variance $\widehat{\sigma}^2 = 7.853$, and Level 2 variance $\widehat{\tau}_{00} = 0.319$. So the intraclass correlation $\widehat{\rho} = \widehat{\tau}_{00} / \left(\widehat{\tau}_{00} + \widehat{\sigma}^2\right) = 0.319 / (0.319 + 7.853) = 0.039$, which indicates that 3.9% of the variance in the

**Table 3** Results From ANOVA Model With Homogeneous $\sigma^2$: $Y_{ij} = \beta_{0j} + r_{ij}$; $\beta_{0j} = \gamma_{00} + u_{0j}$

| Effect | | Coefficient | *SE* | *T*-Ratio | *df* | *p* |
|---|---|---|---|---|---|---|
| Fixed | Average admin mean $\gamma_{00}$ | 14.995 | 0.036 | 419.108 | 253 | < .001 |
| | | Variance component | SD | $\chi^2$ | *df* | *p* |
| Random | Admin mean $u_{0j}$ | 0.319 | 0.565 | 13850.5 | 253 | < .001 |
| | Level 1 effect $r_{ij}$ | 7.853 | 2.802 | | | |

*Note*. Admin = administration; reliability estimate of random Level 1 coefficient $\beta_{0i} = 0.977$.

**Table 4** Results From ANOVA Model With Heterogeneous $\sigma^2$: $Y_{ij} = \beta_{0j} + r_{ij}$; $\beta_{0j} = \gamma_{00} + u_{0j}$; $\ln(\sigma_{ij}^2) = \alpha_0 + \alpha_1$Edu1 $+ \alpha_2$Occ3 $+ \alpha_3$Stu $+ \alpha_4$Ove $+ \alpha_5$Tim1 $+ \alpha_6$Pur1

| Effect | | Coefficient | *SE* | *T*-Ratio | *df* | *p* |
|---|---|---|---|---|---|---|
| Fixed | Average admin mean $\gamma_{00}$ | 14.972 | 0.035 | 432.435 | 253 | <.001 |
| | | Variance component | *SD* | $\chi^2$ | *df* | *p* |
| Random | Admin mean $u_{0j}$ | 0.297 | 0.545 | 13753.304 | 253 | <.001 |
| | Level-1 variance model | Coefficient | *SE* | *Z*-Ratio | | *p* |
| | Intercept $\alpha_0$ | 2.681 | 0.013 | 203.039 | | <.001 |
| | Edu1 $\alpha_1$ | −0.238 | 0.007 | −34.669 | | <.001 |
| | Occ3 $\alpha_2$ | −0.158 | 0.005 | −29.043 | | <.001 |
| | Stu $\alpha_3$ | −0.096 | 0.003 | −31.206 | | <.001 |
| | Ove $\alpha_4$ | 0.099 | 0.002 | 45.714 | | <.001 |
| | Tim1 $\alpha_5$ | −0.137 | 0.006 | −23.840 | | <.001 |
| | Pur1 $\alpha_6$ | −0.293 | 0.006 | −50.986 | | <.001 |

*Note*. Admin = administration; Edu = education; Occ = occupation; Stu = study; Ove = overseas; Tim = time; Pur = purpose.

test scores was between-administration. Therefore, most score variation came from within-administration. The low intraclass correlation suggests a lower degree of dependence of examinees' scores within each administration. However, the between-administration variance was still significantly larger than 0, with $\chi^2 = 13850.494$, $df = 253$, and $p < 0.001$. From the variance of administration means $\widehat{\tau}_{00} = 0.319$, we expect 95% of the administration means falls within the range $14.995 \pm 1.96 * \sqrt{0.319} = (13.888,\ 16.102)$. If we use the criterion in Shewhart 3-sigma control charts, the lower control limit (LCL) will be 13.301, and the upper control limit (UCL) will be 16.689, which is close to the observed score mean range, that is, 13.29 – 16.76. Therefore, psychometrically, the score means fluctuated across administrations, and the between-administration variance should not be ignored in analyses.

These results were all based on the ANOVA model with the assumption of homogeneity of Level 1 variance $\sigma^2$. The likelihood ratio test ($\chi^2 = 5539.728$, $df = 253$, $p < 0.001$) suggests that $\sigma^2$ was not homogeneous across administrations. Although the estimation of fixed effects and their standard errors is robust to violation of this assumption (Kasim & Raudenbush, 1998), given the big range of score variances across administrations (i.e., $2.15^2$ – $3.61^2$ or 4.62 – 13.03, see Table 2), the heterogeneity of $\sigma^2$ was modeled as a function of six level-1 background variables after evaluating different models for $\sigma^2$

$$\ln\left(\sigma_{ij}^2\right) = \alpha_0 + \alpha_1\text{Edu1} + \alpha_2\text{Occ3} + \alpha_3\text{Stu} + \alpha_4\text{Ove} + \alpha_5\text{Tim1} + \alpha_6\text{Pur1}.$$

Based on results from this model (see Table 4 for the detail), test score variance was related to examinees' background variables. For example, with other background variables controlled, examinees with *undergraduate* education level (Edu1), *full-time students* (Occ3), examinees having taken the test *once* (Tim1), and examinees taking the test for *job application* purpose (Pur1) tended to be more homogeneous; the variance of examinees' scores appeared to decrease with *English study time* (Stu).

Model comparison of homogeneous and heterogeneous Level 1 variance models for examinees' test scores ($\chi^2_{\text{difference}} = 12056.882$, $df_{\text{difference}} = 6$, $p < 0.001$) suggests that the model with heterogeneous $\sigma^2$ fit the data much better than the model with homogenous $\sigma^2$. However, the statistical estimates did not change much, with $\widehat{\gamma}_{00} = 14.972$, $\widehat{\lambda} = 0.977$, $\widehat{\tau}_{00} = 0.297$. With heterogeneous $\sigma^2$ in the model, the Level 1 variance estimate was not available as a single measure, so the

intraclass correlation and proportion variance from within- and between- administration could not be estimated. With the consideration of both model fit and variance estimation, the following analyses would include heterogeneous $\sigma^2$ as part of different multilevel models, unless Level 1 variance estimation was highly demanded (e.g., $\sigma^2$ was needed to estimate variance explained by Level 1 predictors) or the heterogeneity of $\sigma^2$ substantially dropped with predictors included in Level 1 models. The results from the ANOVA model, especially $\sigma^2$ and $\widehat{\tau}_{00}$, would be used as base statistics to evaluate the predictive power of different models.

## Predicting Test Performance at 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 at the administration level. This type of model was first used separately for each group composition variable, then used for the combination of those variables to predict administration means.

Table 5 summarizes results from different means-as-outcome regression models. Based on $\widehat{\gamma}_{00}$ estimates, for each group composition variable, after the percentage of a certain subgroup (e.g., examinees with *undergraduate* education level, *full-time employed* examinees) was controlled at the average level across all administrations (note that Level 2 predictors were centered around their grand means), the grand mean estimate $\widehat{\gamma}_{00}$ was all close to the original grand mean $\widehat{\gamma}_{00} = 14.972$ from the ANOVA model. However, based on $\widehat{\gamma}_{01}$ estimates, each selected group composition variable had a statistically significant relationship with administration means, except the one based on *education level* ($p = 0.430$). In addition, the administration means $\widehat{\beta}_{0j}$ would increase or decrease at different degrees for the same percent change of different subgroups. For example, when the group composition based on *education level* (i.e., percent of *undergraduate* level examinees) increased by 10%, the administration mean would remain almost the same, with a trivial decrease of 0.009 * 10 = 0.09; if the group composition based on *occupation* (i.e., percent of *full-time employed* examinees) increased by 10%, the administration mean would drop by 0.014 * 10 = 0.14; when the group composition based on *overseas English experience* (i.e., the percent of examinees with *0–6 months* of overseas experience) increased by 10%, the administration mean would decrease by 0.052 * 10 = 0.52.

Another way to evaluate the prediction of administration means on group composition is to estimate proportion of variance explained by different group composition variables in the Level 2 model by

$$\frac{\widehat{\tau}_{00}\,(\text{random ANOVA}) - \widehat{\tau}_{00}\,\big(\text{group composition}\big)}{\widehat{\tau}_{00}\,(\text{random ANOVA})},$$

where the $\widehat{\tau}_{00}\,(\text{random ANOVA}) = 0.297$ from the ANOVA model with heterogeneous $\sigma^2$. The column of $\widehat{\tau}_{00}$ in Table 5 shows the variance estimates from different means-as-outcome regression models. The last column shows the proportion of variance of score means explained by different group composition variables. For example, the percent of score means' variance explained by the group composition change in *overseas English experience* was (0.297 – 0.219)/0.297 = 0.263 = 26.3%. From the table, the group composition changes in *education level* and *English use time* could only account for less than 5% of administration means' variance, respectively. However, group changes based on self-evaluated *English difficulty*, *overseas English experience*, and *test-taking times* could separately explain over 20% of the score means' variance.

The results based on single-predictor models were used to select the best combination of predictors for administration means. To avoid collinearity, correlations between different Level 2 predictors were examined. It was found that there were high correlations between group composition variables based on *occupation* and *test-taking purpose* (0.966), between *English difficulty* and *overseas English experience* (0.735), and between *overseas English experience* and *test-taking time* (0.667). So these highly correlated predictors were not simultaneously included in a model. Different models were tried, and the best combined-predictor model included the following group composition variables: *occupation*, *English study time*, *English use time*, and *overseas English experience*. These predictors together explained about 38.7% of the variance in administration means (see the bottom row in Table 5).

## Predicting Test Performance at 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 background variable and then used for the combination of those variables to predict examinees' scores.

**Table 5** Results From Means-as-Outcome Regression Models

| Level 1 model | Level 2 model | $\widehat{\gamma}_{00}$ | $\widehat{\gamma}_{01}\,(p)$ | $\widehat{\tau}_{00}$ | Variance explained % |
|---|---|---|---|---|---|
| $Y_{ij}=\beta_{0j}+r_{ij}$ | $\beta_{0j}=\gamma_{00}+\gamma_{01}\text{Gedu}_j+u_{0j}$ | 14.973 | −0.009 (.430) | 0.296 | 0.3 |
| $Y_{ij}=\beta_{0j}+r_{ij}$ | $\beta_{0j}=\gamma_{00}+\gamma_{01}\text{Gocc}_j+u_{0j}$ | 14.972 | −0.014 (< .001) | 0.274 | 7.7 |
| $Y_{ij}=\beta_{0j}+r_{ij}$ | $\beta_{0j}=\gamma_{00}+\gamma_{01}\text{Gstu}_j+u_{0j}$ | 14.972 | 0.073 (< .001) | 0.258 | 13.1 |
| $Y_{ij}=\beta_{0j}+r_{ij}$ | $\beta_{0j}=\gamma_{00}+\gamma_{01}\text{Guse}_j+u_{0j}$ | 14.973 | −0.033 (.002) | 0.284 | 4.4 |
| $Y_{ij}=\beta_{0j}+r_{ij}$ | $\beta_{0j}=\gamma_{00}+\gamma_{01}\text{Gdif}_j+u_{0j}$ | 14.973 | 0.073 (.002) | 0.228 | 23.2 |
| $Y_{ij}=\beta_{0j}+r_{ij}$ | $\beta_{0j}=\gamma_{00}+\gamma_{01}\text{Gove}_j+u_{0j}$ | 14.973 | −0.052 (.002) | 0.219 | 26.3 |
| $Y_{ij}=\beta_{0j}+r_{ij}$ | $\beta_{0j}=\gamma_{00}+\gamma_{01}\text{Gtim}_j+u_{0j}$ | 14.972 | 0.030 (.002) | 0.231 | 22.2 |
| $Y_{ij}=\beta_{0j}+r_{ij}$ | $\beta_{0j}=\gamma_{00}+\gamma_{01}\text{Gpur}_j+u_{0j}$ | 14.972 | −0.022 (.002) | 0.266 | 10.4 |
| $Y_{ij}=\beta_{0j}+r_{ij}$ | $\beta_{0j}=\gamma_{00}+\gamma_{01}\text{Gocc}_j+\gamma_{02}\text{Gstu}_j+\gamma_{03}\text{Guse}_j+\gamma_{04}\text{Gove}_j+u_{0j}$ | 14.972 | $\widehat{\gamma}_{01}=-0.016$ (< .001)<br>$\widehat{\gamma}_{02}=0.035$ (.018)<br>$\widehat{\gamma}_{03}=-0.038$ (.002)<br>$\widehat{\gamma}_{04}=-0.036$ (< .001) | 0.182 | 38.7 |

*Note*. Gedu = group education; Gocc = group occupation; Gstu = group study; Guse = group use; Gdif = group difficulty; Gove = group overseas; Gtim = group time; Gpur = group purpose.

Model comparisons suggest that after Level 1 predictors entered into the ANOVA model: (a) Level 1 residual variance $\sigma^2$ was still heterogeneous in all models, and (b) the original model used in ANOVA for heterogeneity of $\sigma^2$ remained effective in all the random-coefficient regression models. Therefore, models with heterogeneity of $\sigma^2$ were used to estimate statistics for fixed effects $\widehat{\gamma}$ (i.e., the conditional grand means and average regression coefficients) so that we can examine the relationships of different background variables with examinees' scores.

Another index to examine the prediction power of Level 1 variables is the variance explained by examinees' background information using

$$\frac{\widehat{\sigma}^2\,(\text{random ANOVA})-\widehat{\sigma}^2\,(\text{background})}{\widehat{\sigma}^2\,(\text{random ANOVA})},$$

where the $\widehat{\sigma}^2$ (random ANOVA) = 7.853 from the ANOVA model with homogeneous $\sigma^2$. For each background variable, the random-coefficient regression model with homogenous $\sigma^2$ were also run to have the estimation of the Level 1 variance $\widehat{\sigma}^2$ (background). Then the proportion of variance explained by each background variable was estimated. For each of those models, one $\widehat{\sigma}^2$ might not be a good index for the variety of Level 1 variance across administrations, but it helped us have an approximate estimation of the predictors' contributions. In addition, comparing the results from models with homogeneous and heterogeneous $\sigma^2$ did not find big differences in the estimation of fixed effects.

Tables 6 and 7 show the models, the average intercepts and slopes, and the proportions of variance explained by each of the four categorical and four ordinal background variables. To check the variability of the background-score relationships across administrations, the standard deviations of the intercepts and slopes were also included in the tables. Although all background variables had statistically significant relationships with test scores ($p$ values were not provided in Tables 6 and 7), different background variables explained different proportions of test score variance (see the last column in Tables 6 and 7). The specific relationships of the four categorical background variables with test scores are described below:

- For the predictor *education level*, the examinees with *college and below* levels had the average score of 14.049, and examinees at *undergraduate*, *graduate*, and *language institution* levels tended to have higher scores by 0.976, 0.717, and 0.955.
- For the predictor *occupation*, the *full-time employed* examinees had an average score of 14.260, and the *part-time employed*, *not employed*, and *full-time students* had higher scores by 0.684, 0.796, and 0.894.
- For the predictor *test-taking times*, the first-time test takers had an average score of 14.895, the second- and third-time test takers had slightly higher scores by 0.233 and 0.184, but the fourth- and more time test takers had a slightly lower score by 0.147.
- For the predictor *test-taking purpose*, the test takers for *promotion* purpose had the lowest test score mean of 13.553, and all subgroups with other purposes had higher scores by 1.5 – 1.9.

**Table 6** Results From Random-Coefficient Regression Using Categorical Background Variables

| Level 1 model | Level 2 model | $\hat{\gamma}$ (*SD*) | $\hat{\sigma}^2$ | Variance explained % |
|---|---|---|---|---|
| $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Edu1}_{1ij} + \beta_{2j}\text{Edu2}_{2ij} + \beta_{3j}\text{Edu3}_{3ij} + r_{ij}$ | $\beta_{0j} = \gamma_{00} + u_{0j}$ | 14.049 (.871) | 7.748 | 1.3 |
| | $\beta_{1j} = \gamma_{10} + u_{1j}$ | 0.976 (.669) | | |
| | $\beta_{2j} = \gamma_{20} + u_{2j}$ | 0.717 (.644) | | |
| | $\beta_{3j} = \gamma_{30} + u_{3j}$ | 0.955 (.698) | | |
| $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Occ1}_{1ij} + \beta_{2j}\text{Occ2}_{2ij} + \beta_{3j}\text{Occ3}_{3ij} + r_{ij}$ | $\beta_{0j} = \gamma_{00} + u_{0j}$ | 14.260 (.559) | 7.654 | 2.5 |
| | $\beta_{1j} = \gamma_{10} + u_{1j}$ | 0.684 (.603) | | |
| | $\beta_{2j} = \gamma_{20} + u_{2j}$ | 0.796 (.535) | | |
| | $\beta_{3j} = \gamma_{30} + u_{3j}$ | 0.894 (.433) | | |
| $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Tim1}_{1ij} + \beta_{2j}\text{Tim2}_{2ij} + \beta_{3j}\text{Tim3}_{3ij} + r_{ij}$ | $\beta_{0j} = \gamma_{00} + u_{0j}$ | 14.895 (.590) | 7.806 | 0.6 |
| | $\beta_{1j} = \gamma_{10} + u_{1j}$ | 0.233 (.218) | | |
| | $\beta_{2j} = \gamma_{20} + u_{2j}$ | 0.184 (.310) | | |
| | $\beta_{3j} = \gamma_{30} + u_{3j}$ | −0.147 (.460) | | |
| $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Pur1}_{1ij} + \beta_{2j}\text{Pur2}_{2ij} + \beta_{3j}\text{Pur3}_{3ij} + \beta_{4j}\text{Pur4}_{4ij} + r_{ij}$ | $\beta_{0j} = \gamma_{00} + u_{0j}$ | 13.553 (.503) | 7.543 | 4.0 |
| | $\beta_{1j} = \gamma_{10} + u_{1j}$ | 1.547 (.476) | | |
| | $\beta_{2j} = \gamma_{20} + u_{2j}$ | 1.864 (.434) | | |
| | $\beta_{3j} = \gamma_{30} + u_{3j}$ | 1.559 (.394) | | |
| | $\beta_{4j} = \gamma_{40} + u_{4j}$ | 1.737 (.635) | | |

*Note*. $\hat{\gamma}$ is the average of intercepts or slopes across administrations; *SD* is the standard deviation of the intercepts or slopes across administrations. Edu = education; Occ = occupation; Tim = time; Pur = purpose.

**Table 7** Results From Random-Coefficient Regression Using Ordinal Background Variables

| Level 1 model | Level 2 model | $\hat{\gamma}$ (*SD*) | $\hat{\sigma}^2$ | Variance explained % |
|---|---|---|---|---|
| $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Stu}_{ij} + r_{ij}$ | $\beta_{0j} = \gamma_{00} + u_{0j}$ | 13.484 (.771) | 7.711 | 1.8 |
| | $\beta_{1j} = \gamma_{10} + u_{1j}$ | 0.427 (.113) | | |
| $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Use}_{ij} + r_{ij}$ | $\beta_{0j} = \gamma_{00} + u_{0j}$ | 13.775 (.620) | 7.624 | 2.9 |
| | $\beta_{1j} = \gamma_{10} + u_{1j}$ | 0.446 (.104) | | |
| $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Dif}_{ij} + r_{ij}$ | $\beta_{0j} = \gamma_{00} + u_{0j}$ | 17.934 (.682) | 7.208 | 8.2 |
| | $\beta_{1j} = \gamma_{10} + u_{1j}$ | −0.865 (.119) | | |
| $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Ove}_{ij} + r_{ij}$ | $\beta_{0j} = \gamma_{00} + u_{0j}$ | 13.116 (.540) | 6.608 | 15.9 |
| | $\beta_{1j} = \gamma_{10} + u_{1j}$ | 0.946 (.107) | | |
| $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Stu}_{1ij} + \beta_{2j}\text{Use}_{2ij} + \beta_{3j}\text{Dif}_{3ij} + \beta_{4j}\text{Ove}_{4ij} + r_{ij}$ | $\beta_{0j} = \gamma_{00} + u_{0j}$ | 13.518 (.730) | 6.185 | 21.2 |
| | $\beta_{1j} = \gamma_{10} + u_{1j}$ | 0.362 (.092) | | |
| | $\beta_{2j} = \gamma_{20} + u_{2j}$ | 0.213 (.078) | | |
| | $\beta_{3j} = \gamma_{30} + u_{3j}$ | −0.570 (.073) | | |
| | $\beta_{4j} = \gamma_{40} + u_{4j}$ | 0.778 (.092) | | |

*Note*. $\hat{\gamma}$ is the average of intercepts or slopes across administrations; *SD* is the standard deviation of the intercepts or slopes across administrations. Stu = study; Dif = difficulty; Ove = overseas.

The average slope estimates $\hat{\gamma}$ in Table 7 show the relations of the four ordinal background variables with test scores. For example, with one unit of scale increase in *English study time*, *English use time*, and *overseas English experience*, on average, the examinees' scores tended to increase by 0.427, 0.446, and 0.946, respectively. One unit increase of *English difficulty* resulted in a score decrease by 0.865.

However, these relationships of examinees' background variables with their test scores were based on average estimates across administrations. A closer look at the random effects of all random-coefficient models found that all slopes' variances were statistically significant ($p \leq 0.001$, not provided in Tables 6 and 7), which means that the relationships between examinees' test scores and their background varied significantly across administrations. The standard deviations of the slopes based on these models (see Tables 6 and 7) show specifically in what extent the relationships varied across

administrations. From the column $\hat{\gamma}$ (*SD*) in Table 6, for most categorical background variables, their associations with test scores were so different across administrations that the directions of the associations (i.e., the sign of slopes) might be opposite in different administrations (see the big standard deviations compared with the average slope estimates per se). For example, for the relationship of *occupation* with test scores, 95% of the slopes would be in the range of $0.684 \pm 1.96 * 0.603 = (-0.498, 1.866)$ for $\hat{\gamma}_{10}$, $0.796 \pm 1.96 * 0.535 = (-0.253, 1.845)$ for $\hat{\gamma}_{20}$, and $0.894 \pm 1.96 * 0.433 = (0.045, 1.743)$ for $\hat{\gamma}_{30}$. The variability of relationships was relatively smaller for the background based on *test-taking purpose*. Based on the column $\hat{\gamma}$ (*SD*) in Table 7, the associations of all ordinal background variables with test scores were relatively more consistent and in the same direction across administrations (see the much smaller standard deviations compared with the average slope estimates).

To explore how those score-background relationships changed across administrations, some Level 2 predictors were included to model Level 1 intercepts and slopes (i.e., intercepts- and slopes-as-outcomes models) in follow-up analyses (see Tables A1 to A10 in the Appendix). The results from those models show the following patterns:

- The *education level*-score relationships did not change with the group composition based on *overseas English experience* (see $\gamma_{11}$, $\gamma_{21}$, and $\gamma_{31}$ in Table A1), but did change with the group composition based on *English use time* (see $\gamma_{11}$, $\gamma_{21}$, and $\gamma_{31}$ in Table A3): the *education level*-score relationships became weak when there were more examinees who used English *less than 20% of time* in their daily life in the administration.
- The *occupation*-score relationships changed with the group composition based on *overseas English experience* (Table A5).
- The *English study time*-score relationships varied with the group composition based on *English use time* (Table A7).
- The *English use time*-score relationships changed with group composition based on *occupation* (Table A9).

In addition, all slopes had negative relationships with intercepts (see Tables A2, A4, A6, A8, and A10), which indicates that when the administration means were higher, the relationships of background with test scores would be smaller.

To choose a good combination of Level 1 predictors for test scores, the proportion of variance explained by different background variables needs to be evaluated. From the last column in Tables 6 and 7, the categorical background variable *test-taking purpose* and most ordinal background variables had bigger contributions to scores' variance. For the sake of parsimony in statistical modeling and the requirement of fewer variables for ordinal predictors, the four ordinal background variables were selected for the model. In addition, the categorical variables had more random effects, and no high correlations were found among the four ordinal predictors. Therefore, the four ordinal background variables were included in the final model to predict test scores (see the bottom panel in Table 7), and no significant improvement was made by adding more categorical background variables. Based on results from this model, all four selected background variables had statistically significant relationships with test scores and could explain 21% of the test-score variance. Although the variances of intercepts and slopes were statistically significant (not provided in Table 7), the association of all ordinal background variables with test scores was relatively consistent across administrations (see the smaller standard deviations compared with the average slope estimates in the bottom panel in Table 7).

## Using Background Information to Predict Test Performance: Validation

Based on the best random-coefficient regression model (see the bottom panel in Table 7), at the individual examinee level, about 21% of score variance could be explained by four background variables and the prediction error (i.e., root mean squared error or RMSE) was $\sqrt{6.185} = 2.49$. Compared to the standard deviation (i.e., 2.86) of examinees' observed test scores, the prediction error was just slightly smaller, and the prediction was not strong. Based on the best regression with means-as-outcomes model (see the bottom panel in Table 5), about 39% of the means' variance could be accounted for by four group composition variables, with RMSE $= \sqrt{0.182} = 0.43$. Compared with standard deviation (i.e., 0.57) of the observed score means, the prediction error was smaller, and the prediction power was better. So we used this regression with means-as-outcomes model to predict 22 new administrations' means and then compared them with operational scoring results. Table 8 shows the predicted means, operational means, and their differences. Compared with operational means, the predicted means were higher on some administrations but lower on the other administrations. The mean differences varied from −0.42 to 0.81, and about 82% (i.e., 18/22) of the absolute values of mean differences were smaller than the RMSE (i.e., 0.43) estimated in the prediction model. This is consistent with the finding from the model that 68% of the actual means should be within ±0.43 range of the predicted means. Although the differences varied across

**Table 8** New Administrations' Predicted and Operational Means

| Administration | Predicted mean | Operational mean | Mean difference |
|---|---|---|---|
| 1 | 14.74 | 14.66 | −0.08 |
| 2 | 14.75 | 14.64 | −0.11 |
| 3 | 14.48 | 14.27 | −0.21 |
| 4 | 14.79 | 14.65 | −0.14 |
| 5 | 14.73 | 14.61 | −0.12 |
| 6 | 14.75 | 14.55 | −0.20 |
| 7 | 14.67 | 15.20 | 0.53 |
| 8 | 14.77 | 14.72 | −0.05 |
| 9 | 14.39 | 14.28 | −0.11 |
| 10 | 14.87 | 14.76 | −0.11 |
| 11 | 14.91 | 14.49 | −0.42 |
| 12 | 14.65 | 15.14 | 0.49 |
| 13 | 14.78 | 15.18 | 0.40 |
| 14 | 14.56 | 14.65 | 0.09 |
| 15 | 14.51 | 14.68 | 0.17 |
| 16 | 14.58 | 14.90 | 0.32 |
| 17 | 14.58 | 14.85 | 0.27 |
| 18 | 14.34 | 14.88 | 0.54 |
| 19 | 14.35 | 15.16 | 0.81 |
| 20 | 14.58 | 14.79 | 0.21 |
| 21 | 14.97 | 15.08 | 0.11 |
| 22 | 14.09 | 13.69 | −0.40 |
| Mean | 14.63 | 14.72 | 0.09 |
| *SD* | 0.21 | 0.35 | 0.33 |
| Minimum | 14.09 | 13.69 | −0.42 |
| Maximum | 14.97 | 15.20 | 0.81 |

administrations, the average differences were very small (i.e., 0.09). Therefore, the prediction model for score means was confirmed and validated based on the 22 new administrations' data.

## Discussion

Different methods and techniques have been proposed to monitor test scores across administrations (von Davier, 2012). The relationships of examinees' background with their test performance have also been explored in some studies for the quality control purpose (Lee et al., 2013; Luo et al., 2011; Wei, 2013). In this study, the test data collected from 254 administrations were considered as having a two-level hierarchical structure, with examinees at Level 1 and administrations at Level 2. The multilevel analysis model was used to explore the relationships between examinees' background and their scores at both levels.

The one-way ANOVA with random effects model provided basic descriptive information of the test scores. Based on the model, the grand mean estimate was very precise, with 95% confidence interval of (14.924, 15.066), and the group mean estimates were very reliable, with the general reliability estimate of 0.977. The intraclass correlation of 3.9% suggests that most score variance came from within-administration, and there was a low degree of dependence of scores within administrations. The lower proportion of between-administration variance does not indicate that we can ignore it because the score means fluctuated across administrations, with 95% confidence interval of (13.888, 16.102). This may be a concern from the test quality-control perspective. Why did the score means vary significantly across administrations? Were test forms equivalent across administrations? Were the scoring rubrics changed over administrations? Was the raters' scoring performance consistent over time? Were there any population changes across administrations? Understanding and monitoring score means' fluctuation over administrations is especially important for a testing program with no equating due to the small number of the easy-to-remember constructed response items in the test. The quality control can be conducted from different perspectives using different studies and procedures. The multilevel analyses in this study first focused on exploring the relationships between means fluctuation and population change across administrations.

Based on the regression with means-as-outcomes model, all selected group composition variables had strong relationships with group means, except the one based on *education level*. Three variables could separately account for 22% to 26% of means' variance, two variables could separately account for 10% to 13% of variance, and two variables could explain less than 10% of variance. The best single predictor for score means was the group composition based on *overseas English experience*, which could explain 26% of means' variance. The best combined four group-composition variables were based on *occupation*, *English study time*, *English use time*, and *overseas English experience*; they could together account for 39% of the means' variance, with a prediction error of 0.43. Given the range of observed score means of 13.29 – 16.76 and standard deviation of 0.57, the group-composition variables had a fairly powerful prediction for administration means. The validity of the prediction model was confirmed by comparing predicted means and operational means for 22 new administrations.

The results from this study provided a strong empirical support for the hypothesis proposed but not proved in the study by Luo et al. (2011), that is, "the variation in the examinee composition across administrations is a major reason for fluctuations in the mean of the scaled scores" (p. 2). However, 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. (2011).

Compared with the results at Level 2, the prediction of individuals' scores based on their background variables at Level 1 was relatively weak, although all selected background variables separately or collectively had statistically significant relationships with test scores. Based on the random-coefficient regression model, seven out of eight background variables could separately explain less than 10% of scores' variance. Consistent with what we found at Level 2, the best single predictor for test scores was *overseas English experience*, which could explain 16% of scores' variance. Therefore, the background information based on *overseas English experience* was the best single predictor at both the examinee and administration levels. The best combined four background variables for test scores were *English study time*, *English use time*, *English difficulty*, and overseas English experience, and they could together account for 21% of the test scores' variance, with a prediction error of 2.487. The range of observed scores was from 0 to 24, with a standard deviation of 2.86. So examinees' background variables had a weak prediction for their test scores.

The finding of a strong test performance prediction at the administration level and a weak prediction at the examinee level is not surprising, given that different types of variables and units of analysis were used at the two levels. The Level 1 predictors were basically categorical variables, and Level 2 predictors were continuous 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 was also found in another study (Wei, 2013), which used examinees' background information to predict their English listening and reading performance.

The multilevel analysis has strength in evaluating both fixed and random effects in prediction models. The significant random effects of both intercepts and slopes (see the standard deviations of the intercepts and slopes in Tables 6 and 7) suggests that the score-background relationships, both the direction (i.e., the sign of slopes) and the strength of those relationships (i.e., the value of slopes), varied across administrations. In other words, the subgroups' performance and their difference changed with administrations. In operational work, it is not unusual to use subgroups' performance in previous administrations and subgroups' sample sizes in current administration to predict current test performance. The finding from this study indicates that, at least for this test, it may be not appropriate to weight subgroups' average scores by their frequencies to predict or verify an administration's test performance. The same conclusion was made in the study for an English listening and reading test program (Wei, 2013).

Not only can multilevel analysis estimate the variability of score-background relationships across administrations, it can also show us how these relationships change across administrations by the intercepts- and slopes-as-outcomes model. Based on the results from the intercepts- and slopes-as-outcomes models, the association of test scores with examinees' background might depend on group composition in administrations. For example, the relationships between *education level* and test scores and between *English study time* and test scores varied with group composition based on *English use time* in administrations; the *occupation*-score relationships changed with group composition based on *overseas English experience*; the relationships between *English use time* and test scores depended on the group composition based on *occupation*. In addition, the score-background association might also depend on group proficiency level in administrations.

For example, the negative correlations between intercepts and slopes indicate that the score-background relationships tended to be stronger when the test score means or conditional means were lower.

The selection of models for quality control in operational work depends on the specific purposes. The one-way ANOVA with random effects model is useful in estimating the population score mean, the administration means' range and reliability, the within- and between-administration score variances, and score dependence within administrations. If the purpose is to predict examinees' test scores based on their characteristics or their scores from other tests, the random-coefficient model may be used. If the purpose is to predict or verify administration score means, we can use the regression model with means-as-outcomes. If we want to evaluate in detail how the relationships between examinees' test scores and their characteristics change across administrations, the intercepts- and slopes-as-outcomes model should be the choice.

The application of the prediction model in quality control depends on how strong the relationship is and how powerful the prediction is. For example, the best prediction model for administration means in this study had $R^2$ of 0.39, and the correlation between score means and group composition was $\sqrt{0.39} = 0.62$. So the model can be used to understand the fluctuation of administration means, especially the unusually high or low means, through examinees' group composition change. In another study for an English listening and reading test (Wei, 2013), the best prediction model for administration means had $R^2$ of 0.85, and the correlation between score means and group composition was 0.92. So the model can be used to monitor test score means by evaluating different equating results based on examinees' background information.

## Conclusion

Based on the multilevel analysis of 330,091 examinees' test scores and background information collected from 254 administrations of an English-speaking test, this study found: (a) at the examinee level, the examinees' background information had statistically significant relationships with their test scores, and the relationships varied across administrations; however, the prediction of individuals' test scores based on their background variables was not strong, and (b) at the administration level, the group composition had statistically significant relationships with administration means; the prediction of administration means based on group composition variables was fairly strong, and the model had potential applications in understanding and monitoring test performance across administrations. The results suggest that multilevel analysis has potential in evaluating test performance across administrations by exploring and applying the relationships between examinees' characteristics and their test performance at both individual and administration levels.

## Future Research

The study was to explore the potential of multilevel analysis in examining and evaluating the relationship between examinees' test performance and their background information. The finding from this study is positive and promising. However, there are some limitations in this study, and future research may address the following issues.

First, the data in this study were collected from a performance assessment with a small number of constructed response items. In addition, the examinees' background information was the only type of variables used to predict the test performance. For the dependent variables, the comparability of test scores was controlled by using consistent test development procedures and scoring rubrics; for the independent variables, the background information was represented by categorical and ordinal variables at Level 1. This may have attenuated their relationships in statistical modeling and estimation.

Second, it is well known that while using regression models to predict criterion variables, there is the phenomenon of *regression toward the mean* effect as long as there is a less than perfect correlation between the criterion variable and predictor(s). In this study, the correlation between group composition and score means was 0.62. We need to be careful while interpreting the predicted test performance for those administrations with very high or low score means.

Third, for testing programs with frequent administrations, examinees' background variables and their relationships with test performance may gradually change. Therefore, it is necessary to reexamine their relationships and adjust the prediction models during the long life of a testing program.

Fourth, in this study, 46% of examinees reported that they had taken the test once, twice, or three and more times while registering to take the test. Although we could not track who attended which administrations, test scores from the same examinees might have been included in the data. Therefore, it is very possible to have some score dependence between different administrations. Future research needs to use data that do not include repeaters' scores or consider the score dependence in the model.

Finally, to explore fully the potential of using multilevel analysis in understanding and monitoring the performance of a test, different types of variables need to be collected and included in the prediction model, such as information about test forms, administrations, equating chains, examinees, and rater pools. The data collected from the administration of a long test with an objective scoring method and good equating design are highly demanded for the research (e.g., Wei, 2013). With more variables included in the multi-administration hierarchical test data, the potential and strength of the multilevel analysis can be further investigated and evaluated.

## Acknowledgments

The authors thank Hongwen Guo, Yi-Hsuan Lee, and Rebecca Zwick for their comments and suggestions on the earlier versions of the paper. The authors also thank Kate Costanzo for providing data for this study.

## References

Adams, R. J., Wilson, M. R., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

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

Boeck, P. D., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Dorans, N. (Ed.). (2004). Assessing the population sensitivity of equating functions [Special issue]. *Journal of Educational Measurement*, *41*(1).

Haberman, S., Guo, H., Liu, J., & Dorans, N. (2008). *Consistency of SAT® Reasoning score conversions* (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. J. (2013). Harmonic regression and scale stability. *Psychometrika*, *78*(4), 815–829.

Lee, Y.-H., Liu, M., & von Davier, A. A. (2013). Detection of unusual test administrations using a linear mixed effects model. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Proceedings of the 77th international meeting of the Psychometric Society* (pp. 133–150). New York, NY: Springer.

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*(3), 557–575.

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.

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.

Qian, J., von Davier, A. A., & Jiang, Y. (2013). Achieving a stable scale for an assessment with multiple forms — Weighting test samples in IRT linking and equating. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Proceedings of the 77th international meeting of the Psychometric Society* (pp. 171–186). New York, NY: Springer.

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 [Computer software]. Lincolnwood, IL: Scientific Software International.

Rijmen, F., Tuerlinckx, F., Boeck, P. D., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205.

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.

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.

## Appendix
## Results From Intercepts and Slopes-as-Outcomes Models

**Table ??** Results From Model: $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Edu1}_{1ij} + \beta_{2j}\text{Edu2}_{2ij} + \beta_{3j}\text{Edu3}_{3ij} + r_{ij}$; $\beta_{0j} = \gamma_{00} + \gamma_{01}\text{Gocc}_j + \gamma_{02}\text{Gove}_j + u_{0j}$, $\beta_{1j} = \gamma_{10} + \gamma_{11}\text{Gove}_j + u_{1j}$, $\beta_{2j} = \gamma_{20} + \gamma_{21}\text{Gove}_j + u_{2j}$, $\beta_{3j} = \gamma_{30} + \gamma_{31}\text{Gove}_j + u_{3j}$

| Effect | | Coefficient | *SE* | *T*-ratio | *df* | *p* |
|---|---|---|---|---|---|---|
| Fixed | For admin means | | | | | |
| | Intercept 1 $\gamma_{00}$ | 14.051 | 0.049 | 286.236 | 251 | <.001 |
| | Occupation $\gamma_{01}$ | −0.019 | 0.002 | −8.117 | 251 | <.001 |
| | Overseas $\gamma_{02}$ | −0.052 | 0.010 | −5.201 | 251 | <.001 |
| | For education Level 2 slope | | | | | |
| | Intercept $\gamma_{10}$ | 0.974 | 0.050 | 19.490 | 252 | <.001 |
| | Overseas $\gamma_{11}$ | 0.001 | 0.010 | 0.063 | 252 | .950 |
| | For education Level 3 slope | | | | | |
| | Intercept $\gamma_{20}$ | 0.715 | 0.051 | 13.948 | 252 | < .001 |
| | Overseas $\gamma_{21}$ | −0.003 | 0.010 | −0.264 | 252 | .792 |
| | For education Level 4 slope | | | | | |
| | Intercept $\gamma_{30}$ | 0.953 | 0.077 | 12.416 | 252 | <.001 |
| | Overseas $\gamma_{31}$ | 0.011 | 0.015 | 0.707 | 252 | .480 |
| | | Variance component | *SD* | $\chi^2$ | *df* | *p* |
| Random | Admin mean $u_{0j}$ | 0.434 | 0.659 | 1038.667 | 251 | <.001 |
| | $u_{1j}$ | 0.443 | 0.665 | 1069.951 | 252 | <.001 |
| | $u_{2j}$ | 0.409 | 0.639 | 750.047 | 252 | <.001 |
| | $u_{3j}$ | 0.479 | 0.692 | 328.884 | 252 | .001 |
| | Level 1 effect $r_{ij}$ | Heterogeneous | | | | |

**Table ??** Reliability and Tau (as Correlations) for Table A1

| | $\beta_{0j}$ (.713) | $\beta_{1j}$(.708) | $\beta_{2j}$(.615) | $\beta_{3j}$(.288) |
|---|---|---|---|---|
| $\beta_{0j}$ | 1 | | | |
| $\beta_{1j}$ | −.757 | 1 | | |
| $\beta_{2j}$ | −.706 | .953 | 1 | |
| $\beta_{3j}$ | −.548 | .839 | .906 | 1 |

**Table ??** Results From Model: $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Edu1}_{1ij} + \beta_{2j}\text{Edu2}_{2ij} + \beta_{3j}\text{Edu3}_{3ij} + r_{ij}$; $\beta_{0j} = \gamma_{00} + \gamma_{01}\text{Gocc}_j + \gamma_{02}\text{Gove}_j + u_{0j}$, $\beta_{1j} = \gamma_{10} + \gamma_{11}\text{Guse}_j + u_{1j}$, $\beta_{2j} = \gamma_{20} + \gamma_{21}\text{Guse}_j + u_{2j}$, $\beta_{3j} = \gamma_{30} + \gamma_{31}\text{Guse}_j + u_{3j}$

| Effect | | Coefficient | *SE* | *T*-ratio | *df* | *p* |
|---|---|---|---|---|---|---|
| Fixed | For admin means | | | | | |
| | Intercept 1 $\gamma_{00}$ | 14.052 | 0.046 | 302.596 | 251 | <.001 |
| | Occupation $\gamma_{01}$ | −0.026 | 0.003 | −9.693 | 251 | <.001 |
| | Overseas $\gamma_{02}$ | −0.041 | 0.007 | −5.766 | 251 | <.001 |
| | For education Level 2 slope | | | | | |
| | Intercept $\gamma_{10}$ | 0.972 | 0.049 | 19.965 | 252 | <.001 |
| | English use $\gamma_{11}$ | −0.047 | 0.010 | −4.489 | 252 | <.001 |
| | For education Level 3 slope | | | | | |
| | Intercept $\gamma_{20}$ | 0.721 | 0.047 | 15.181 | 252 | <.001 |
| | English use $\gamma_{21}$ | −0.084 | 0.012 | −6.868 | 252 | <.001 |

**Table ??** Continued.

| Effect | | Coefficient | *SE* | *T*-ratio | *df* | *p* |
|---|---|---|---|---|---|---|
| | For education Level 4 slope | | | | | |
| | Intercept $\gamma_{30}$ | 0.949 | 0.075 | 12.590 | 252 | <.001 |
| | English use $\gamma_{31}$ | −0.063 | 0.022 | −2.869 | 252 | .005 |
| | | Variance component | *SD* | $\chi^2$ | *df* | *p* |
| Random | Admin mean $u_{0j}$ | 0.379 | 0.616 | 930.764 | 251 | <.001 |
| | $u_{1j}$ | 0.421 | 0.649 | 1033.830 | 252 | <.001 |
| | $u_{2j}$ | 0.331 | 0.575 | 646.751 | 252 | <.001 |
| | $u_{3j}$ | 0.435 | 0.659 | 332.311 | 252 | .001 |
| | Level 1 effect $r_{ij}$ | *Heterogeneous* | | | | |

**Table ??** ***Reliability and Tau (as Correlations) for Table A3***

| | $\beta_{0j}$ (.686) | $\beta_{1j}$(.698) | $\beta_{2j}$(.566) | $\beta_{3j}$(.270) |
|---|---|---|---|---|
| $\beta_{0j}$ | 1 | | | |
| $\beta_{1j}$ | −.734 | 1 | | |
| $\beta_{2j}$ | −.726 | .967 | 1 | |
| $\beta_{3j}$ | −.552 | .832 | .885 | 1 |

**Table ??** ***Results From Model:*** $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Occ1}_{1ij} + \beta_{2j}\text{Occ2}_{2ij} + \beta_{3j}\text{Occ3}_{3ij} + r_{ij}$; $\beta_{0j} = \gamma_{00} + \gamma_{01}\text{Gocc}_j + \gamma_{02}\text{Gove}_j + u_{0j}$, $\beta_{1j} = \gamma_{10} + \gamma_{11}\text{Gove}_j + u_{1j}$, $\beta_{2j} = \gamma_{20} + \gamma_{21}\text{Gove}_j + u_{2j}$, $\beta_{3j} = \gamma_{30} + \gamma_{31}\text{Gove}_j + u_{3j}$

| Effect | | Coefficient | *SE* | *T*-ratio | *df* | *p* |
|---|---|---|---|---|---|---|
| Fixed | For admin means | | | | | |
| | Intercept 1 $\gamma_{00}$ | 14.261 | 0.033 | 435.956 | 251 | <.001 |
| | Occupation $\gamma_{01}$ | −0.012 | 0.002 | −5.048 | 251 | <.001 |
| | Overseas $\gamma_{02}$ | −0.028 | 0.008 | −3.691 | 251 | <.001 |
| | For part-time employed slope | | | | | |
| | Intercept $\gamma_{10}$ | 0.680 | 0.044 | 15.343 | 252 | <.001 |
| | Overseas $\gamma_{11}$ | −0.046 | 0.008 | −5.805 | 252 | <.001 |
| | For unemployed slope | | | | | |
| | Intercept $\gamma_{20}$ | 0.795 | 0.037 | 21.298 | 252 | <.001 |
| | Overseas $\gamma_{21}$ | −0.030 | 0.007 | −4.256 | 252 | <.001 |
| | For full-time student slope | | | | | |
| | Intercept $\gamma_{30}$ | 0.893 | 0.030 | 30.096 | 252 | <.001 |
| | Overseas $\gamma_{31}$ | −0.025 | 0.005 | −4.824 | 252 | <.001 |
| | | Variance component | *SD* | $\chi^2$ | *df* | *p* |
| Random | Admin mean $u_{0j}$ | 0.228 | 0.478 | 1838.053 | 251 | <.001 |
| | $u_{1j}$ | 0.308 | 0.555 | 726.995 | 252 | <.001 |
| | $u_{2j}$ | 0.258 | 0.508 | 1007.657 | 252 | <.001 |
| | $u_{3j}$ | 0.168 | 0.409 | 1225.559 | 252 | <.001 |
| | Level 1 effect $r_{ij}$ | Heterogeneous | | | | |

**Table ??** ***Reliability and Tau (as Correlations) for Table A5***

| | $\beta_{0j}$ (.838) | $\beta_{1j}$(.598) | $\beta_{2j}$(.723) | $\beta_{3j}$(.754) |
|---|---|---|---|---|
| $\beta_{0j}$ | 1 | | | |
| $\beta_{1j}$ | −.266 | 1 | | |
| $\beta_{2j}$ | −.334 | .915 | 1 | |
| $\beta_{3j}$ | −.450 | .743 | .855 | 1 |

**Table ??** ***Results From Model:*** $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Stu}_{ij} + r_{ij}$; $\beta_{0j} = \gamma_{00} + \gamma_{01}\text{Gocc}_j + \gamma_{02}\text{Gove}_j + u_{0j}$, $\beta_{1j} = \gamma_{10} + \gamma_{11}\text{Guse}_j + u_{1j}$

| Effect | | Coefficient | *SE* | *T*-ratio | *df* | *p* |
|---|---|---|---|---|---|---|
| Fixed | For admin mean | | | | | |
| | Intercept 1 $\gamma_{00}$ | 13.474 | 0.045 | 298.515 | 251 | <.001 |
| | Occupation $\gamma_{01}$ | −0.017 | 0.003 | −6.169 | 251 | <.001 |
| | Overseas $\gamma_{02}$ | −0.039 | 0.008 | −5.017 | 251 | <.001 |
| | For English study slope | | | | | |
| | Intercept $\gamma_{10}$ | 0.430 | 0.009 | 46.621 | 252 | <.001 |
| | English use $\gamma_{11}$ | −0.012 | 0.002 | −5.214 | 252 | <.001 |
| | | Variance component | *SD* | $\chi^2$ | *df* | *p* |
| Random | Admin mean $u_{0j}$ | 0.375 | 0.613 | 1016.001 | 251 | <.001 |
| | $u_{1j}$ | 0.010 | 0.102 | 518.364 | 252 | <.001 |
| | Level 1 effect $r_{ij}$ | Heterogeneous | | | | |

**Table ??** ***Reliability and Tau (as Correlations) for Table A7***

| | $\beta_{0j}$ (.730) | $\beta_{1j}$(.495) |
|---|---|---|
| $\beta_{0j}$ | 1 | |
| $\beta_{1j}$ | −.715 | 1 |

**Table ??** ***Results From Model:*** $Y_{ij} = \beta_{0j} + \beta_{1j}\text{Use}_{ij} + r_{ij}$; $\beta_{0j} = \gamma_{00} + \gamma_{01}\text{Gocc}_j + \gamma_{02}\text{Gove}_j + u_{0j}$, $\beta_{1j} = \gamma_{10} + \gamma_{11}\text{Gocc}_j + u_{1j}$

| Effect | | Coefficient | *SE* | *T*-ratio | *df* | *p* |
|---|---|---|---|---|---|---|
| Fixed | For admin mean | | | | | |
| | Intercept 1 $\gamma_{00}$ | 13.768 | 0.036 | 379.497 | 251 | < .001 |
| | Occupation $\gamma_{01}$ | −0.025 | 0.003 | −8.277 | 251 | < .001 |
| | Overseas $\gamma_{02}$ | −0.051 | 0.008 | −6.550 | 251 | < .001 |
| | For English use slope | | | | | |
| | Intercept $\gamma_{10}$ | 0.448 | 0.008 | 56.867 | 252 | < .001 |
| | Occupation $\gamma_{11}$ | 0.004 | 0.001 | 6.093 | 252 | < .001 |
| | | Variance component | *SD* | $\chi^2$ | *df* | *p* |
| Random | Admin mean $u_{0j}$ | 0.275 | 0.525 | 1559.176 | 251 | < .001 |
| | $u_{1j}$ | 0.009 | 0.095 | 618.0387 | 252 | < .001 |
| | Level 1 effect $r_{ij}$ | Heterogeneous | | | | |

**Table ??** ***Reliability and Tau (as Correlations) for Table A9***

| | $\beta_{0j}$(.823) | $\beta_{1j}$(.570) |
|---|---|---|
| $\beta_{0j}$ | 1 | |
| $\beta_{1j}$ | −.553 | 1 |

## Suggested citation:

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

**Action Editor:** Rebecca Zwick

**Reviewers:** Hongwen Guo and Yi-Hsuan Lee