ETS TOEFL

Research Reports

RR - 77
August 2004

Comparability of TOEFL CBT Writing Prompts for Different Native Language Groups

Yong-Won Lee
Hunter Breland
Eiji Muraki

Comparability of TOEFL CBT Writing Prompts for Different Native Language Groups

Yong-Won Lee and Hunter Breland

ETS, Princeton, NJ

Eiji Muraki

Tohoku University, Japan

RR-04-24

ETS is an Equal Opportunity/Affirmative Action Employer.

Abstract

This study has investigated the comparability of computer-based testing (CBT) writing prompts in the Test of English as a Foreign Language™ (TOEFL®) for examinees of different native language backgrounds. A total of 81 writing prompts introduced from July 1998 through August 2000 were examined using a three-step logistic regression procedure for ordinal items. An English language ability (ELA) variable was created by summing the standardized TOEFL Reading, Listening, and Structure scale scores. This ELA variable was used to match examinees of East Asian (Chinese, Japanese, and Korean) and European (German, French, and Spanish) language groups. Although about one third of the 81 prompts were initially flagged because of statistically significant group effects, the effect sizes were too small for any of those flagged prompts to be classified as having an important group effect.

Key words: Computer-based writing assessment, essay prompt comparability, fairness, polytomous DIF (differential item functioning), native languages, logistic regression, proportional odds-ratio model

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service® (ETS®) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

❖ ❖ ❖

A continuing program of research related to the TOEFL test is carried out in consultation with the TOEFL Committee of Examiners. Its 12 members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, reviews and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Because the studies are specific to the TOEFL test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. Many projects require the cooperation of other institutions, however, particularly those with programs in the teaching of English as a foreign or second language and applied linguistics. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (2003-2004) members of the TOEFL Committee of Examiners are:

Micheline Chalhoub-Deville	University of Iowa
Lyle Bachman	University of California, Los Angeles
Deena Boraie	The American University in Cairo
Catherine Elder	Monash University
Glenn Fulcher	University of Dundee
William Grabe	Northern Arizona University
Keiko Koda	Carnegie Mellon University
Richard Luecht	University of North Carolina at Greensboro
Tim McNamara	The University of Melbourne
James E. Purpura	Teachers College, Columbia University
Terry Santos	Humboldt State University
Richard Young	University of Wisconsin-Madison

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org

Web site: www.ets.org/toefl

Acknowledgements

This research project was funded by the Test of English as a Foreign Language (TOEFL) Research Program at ETS. Several members of the ETS staff and external reviewers in addition to the authors contributed to this project. Robert Kantor at ETS served as a general advisor on a number of matters. Cindy Nguyen and Corinne Reslier provided a TOEFL data set used for the analyses and prepared a data layout for the data set. Ting Lu and Youn-Hee Lim also helped us with data manipulation and analyses. We also would like to thank Brent Bridgman, Phil Everson, Marna Golub-Smith, Richard Luecht, Alan Nicewander, and Don Powers for helpful comments about the earlier manuscript of this report.

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

Computer-based test (CBT) administrations of the Test of English as a Foreign Language™ (TOEFL®) began in the summer of 1998. These administrations included computer-based multiple-choice (MC) tests of reading, listening, structure (a multiple-choice test of grammar and sentence structure), and writing. The prompts for the CBT writing section are selected for each examinee from a pool of prompts in a near-random manner, and as a result all examinees do not receive the same prompt. Since only one essay prompt is administered per examinee in TOEFL CBT, it becomes very important to ensure that each writing prompt is as fair as possible to any subgroup. It is equally important to understand the causes of such group-related effects on essay scores and to introduce procedures to minimize any potential biases. If prompt incomparability is caused by serious prompt bias against a certain subgroup of examinees, it could distort the meaning of the essay score for different examinee subgroups and become a potential threat to test score validity (Sheppard, 1982).

For this reason, a considerable amount of research effort has been directed toward investigating the comparability of TOEFL CBT writing prompts for different test-taker subgroups, such as different gender and response mode (e.g., typed, handwritten) groups (Breland, Muraki, & Lee, 2001; Breland, Muraki, Lee, Najarian, & Beyer, 2000; Gentile, Riazantseva, & Cline, 2001; Wolfe & Manalo, 2001). These studies have found that there might be a small to medium impact of gender and response mode on essay scores and have provided some useful information for subsequent prompt review, revision, and retirement to minimize the potential impact. There is one more important group variable suggested by the American Education Research Association, American Psychological Association, and National Council on Measurement in Education (1999) standards for investigation, but it has yet to be examined fully for the TOEFL CBT writing test: native language backgrounds of test takers. It has been suggested that the native language backgrounds of the examinees could affect their performance on second/foreign language tests (Chen & Henning, 1985; Kim, 2001; Ryan & Bachman, 1992; Sasaki, 1991). Of particular interest has been the question of whether the examinees of non-Indo-European language backgrounds score lower on the EFL/ESL (English as a foreign/second language) tests than do examinees of Indo-European language backgrounds because of the relatively greater dissimilarity between their native languages and the English language (Kim, 2001; Ryan & Bachman, 1992). The impact of examinees' native languages on performance on

second or foreign language tests has long been an important issue for test developers and users (Brown & Iwashita, 1998; Ginther & Stevens, 1998; Hale, Rock, & Jirele, 1989; Hinkel, 2002; Oltman, Stricker, & Barrows, 1988; Swinton & Powers, 1980). Differential item function (DIF) studies in particular have investigated differential performance on the test that could be attributable to the native languages of test takers in the ESL/EFL (English as a Second/Foreign Language) contexts (Angoff, 1989; Chen & Henning, 1985; Kim, 2001; Ryan & Bachman, 1992; Sasaki, 1991). However, the methodology employed in previous studies in language assessment was not directly applicable for TOEFL CBT writing prompts to be investigated in the current study because: (a) item responses investigated in most of these studies were confined to dichotomously-scored multiple-choice (MC) language tests including vocabulary, grammar, listening, and reading comprehension; (b) uniform DIF was the main focus of most of these studies; and (c) all of the studies dealt with situations where the internal matching criterion is usually available for the studied items. Among these studies, Kim's (2001) DIF study on the pre-revision version of the Test of Spoken English™ (TSE®) might be an exception to methodological constraints (a) and (b) above, but, even for her study, an internal matching criterion was available. Clearly, there has been a lack of research on essay tests that are scored polytomously, where reliable, internal matching criterion is not usually available.

The research study reported in this paper attempts to move one step beyond the methodology used in the previous studies and to examine the impact of different native language backgrounds on essay scores. To select the most feasible scheme for investigating group effect in this study, the characteristics of matching criteria, items, and kinds of analyses employed had to be carefully considered at the beginning of this study. Since there was no internal matching criterion, an externally available criterion in the same test battery was used. In addition, an investigation method had to be chosen for the study that could examine both the uniform and nonuniform effects on examinees. The method selected was *logistic regression* of essay scores on the matching variable. A more detailed rationale for the methodology adopted follows.

One of the most important challenges in this project was to find an appropriate variable to use for matching examinees of two different language groups on their English writing ability. It is important to detect instances when examinees of equal ability but from different groups do not have the same probability of success on an item (Angoff, 1993; Hambleton, Swaminathan, & Rogers, 1991). Because the TOEFL CBT writing section is an essay test made up of a single

prompt, no internal matching criterion for writing ability is available (Potenza & Dorans, 1995). The only information available is that provided by the three multiple-choice sections of TOEFL CBT (Listening, Structure, and Reading). For these reasons, a decision was made to create a matching variable by summing the standardized scale scores from the three multiple-choice sections based on a recommendation by Penfield and Lam (2000). The underlying assumption is that, if examinees have high English language ability measured by the three sections of the test as a whole, they should perform well overall on the essays, and vice versa (a more detailed rationale will be provided later in the method section).

A second challenge related to the nature of the essay scores was that examinees' essays for each prompt are scored according to a multipoint scoring rubric by two independent raters. In TOEFL CBT, each examinee's essay is rated by two independent raters on a six-point scoring rubric, and the two raters' ratings are averaged for score-reporting purposes. The essay scores are thus discrete and bounded between 1 and 6. This poses a challenge because most of the methodology has been developed with a focus on dichotomous items. A method was required that could handle polytomously scored items. Various methods have been used for polytomous items, including logistic regression (French & Miller, 1996), polytomous IRT (item response theory) (Muraki, 1999; Wainer, Sireci, & Thissen, 1991), the Mantel and Hanszel technique (Zwick, Donoghue, & Grima, 1993), the SIBTEST procedure (Chang, Mazzeo, & Roussos, 1995), the standardization method (Dorans & Schmitt, 1991), and logistic discriminant function analysis (Miller & Spray, 1993). Among these methods, however, methods requiring an internal criterion (e.g., polytomous IRT and polytomous SIBTEST) were not feasible for this study. Moreover, methods such as the Mantel and Hanszel technique, SIBTEST, and the standardization methods, may not be appropriate for detecting the nonuniform effects (Penfield & Lam, 2000), unless a special modification is made (Clauser & Mazor, 1998).

A third challenge resulted from the need to examine nonuniform as well as uniform effects. There has been a considerable amount of research on the effect of test takers' native language backgrounds on the test dimensionality of the EFL/ESL test or on the different factor structure of the same tests across the different levels of language proficiency (Ginther & Stevens, 1998; Oltman, Stricker, & Barrows, 1988; Swinton & Powers, 1980). Some of these studies have suggested that interpretation of TOEFL section scores depends on the examinees' overall level of proficiency, with more differentiated factors for TOEFL sections for low-scoring examinees,

but less distinct constructs for high-scoring examinees. Since the matching variable is from the MC segments of the Listening, Structure, and Reading sections, an argument can be made that native language backgrounds might have differential impact on essay scores at different levels of English language ability. Both IRT and logistic regression can be very effective in detecting nonuniform DIF, but the IRT method needs an internal matching criterion. A clear advantage of the logistic regression method for nonuniform effect is that it may be as accurate as IRT procedures in determining the nature of the group effect within the flagged item because it is model-based and close to IRT in form, but does not require an internal matching criterion as in IRT (French & Miller, 1996; Swaminathan & Rogers, 1990).

The principal objective of this study was to investigate the comparability of TOEFL CBT writing prompts for examinees of different language backgrounds, with a focus on European (German, French, and Spanish) and East Asian (Chinese, Japanese, and Korean) native language groups as "reference" and "focal" groups, respectively. More specifically, this study is designed to assess the performance of examinees in the East Asian and European language groups after they were matched on English language ability, as defined in this study. Methodologically, this study is primarily concerned with procedures for examining nonuniform and uniform effects of an examinee group variable in an essay test, where examinees take a single essay prompt that is scored polytomously.

Methods

Sample

The sample of data analyzed consisted of TOEFL CBT essay data collected from July 1998 through August 2000 administrations for three European (French, German, and Spanish) and three East Asian (Chinese, Japanese, and Korean) language groups. In total, 262,034 essays written on 87 different topics were included in this study. Six prompts with insufficient data for the focal and reference groups were dropped from the analysis. Of the 254,435 examinees and 81 prompts included in the analysis, a total of 121,494 examinees were native speakers of three European languages and 132,941 were native speakers of three East Asian languages. Among the three European languages, Spanish was the largest group (n = 66,282), followed by French (n = 28,007) and German (n = 27,205). Chinese was the largest group (n = 52,112) among the East Asian languages, followed by Japanese (n = 43,666) and Korean (n = 37,163).

Instruments

Data analyzed included scores on the Reading, Listening, Structure, and Writing subtests of the TOEFL CBT. The Listening and Structure tests are adaptive, the Reading test is linear, and the Writing test score is basically the average of the two reader ratings. More detailed descriptions of the TOEFL CBT section scores are as follows: (a) the TOEFL Reading score is based on a linear multiple-choice test of reading and has a score range from 0 to 30; (b) the TOEFL Listening score is based on an adaptive multiple-choice test of listening comprehension and has a score range of 0 to 30; (c) the TOEFL Structure score is based on an adaptive multiple-choice test of English grammar and sentence structure and has a range of 0 to 13; (d) the TOEFL Writing score is based on two independent readings and holistic ratings of the essay response on a 1 to 6 scale and ranges from 1 to 6 with possibilities of 0.5 intervals (see the *Computer-based TOEFL Score User Guide*, ETS, 1998, for more details about the section score scales). For Writing, it is in general the average of two identical or adjacent scores. If the first two ratings differ by more than one point, however, a third reader is used to adjudicate the score, and the two closest ratings are averaged (see Appendix F for scoring rubrics for the TOEFL CBT writing prompts).

A matching variable named "English language ability" (ELA) score was created by: (a) taking all the examinees who took the same writing prompt between July 1998 through August 2000; (b) standardizing the scale scores of the Reading, Listening, and Structure sections separately based on the total examinee samples for a specific prompt; (c) and summing the standardized scores of the three sections for each examinee. Next, only examinees of the six native language groups of interest were selected and included in the analysis. Since the Structure section and the essay scores contribute to the Structure/Written Expression (SWE) section scale score, one might argue that the structure scores (i.e., structure scale scores without essay scores combined) alone could be a more valid matching variable for the writing ability by definition. However, the structure items may not be measuring the same construct as the essay test (DeMauro, 1992). In addition, the correlation between the essay rating and the structure scale score is not significantly larger than that between the essay score and each of the other two MC section scores. Rather, when the scale scores from each of these three sections were standardized and combined, the correlation between the essay score and the matching criterion was maximized. Thus, a decision was made to create a matching variable by summing the

standardized scale scores from the three multiple-choice sections for each of the prompts. The ELA scores were approximately within the range of –10 to 5, with a mean of 0 and a standard deviation of 2.7. When the expected score curves were drawn for the reference and focal groups (to be explained later in the subsequent section), however, the ELA score range of –10 to 10 was used for the sake of symmetry and convenience.

Data Analysis

Logistic regression analysis (Hosmer & Lemeshow, 1989) has been used mainly to study group effect for *dichotomously scored* test items, and this is done by specifying separate equations for the reference and focal groups of examinees (Swaminathan & Rogers, 1990). French and Miller (1996) demonstrated that this procedure can be extended for polytomous items as well. In this study, one of the three polytomous logistic regression procedures used by French and Miller (1996) is extended further to make it possible to compare the expected score curves for reference and focal groups in the context of TOEFL CBT writing prompt investigation. Logistic regression has also two main advantages over linear regression. The first is that the dependent variable does not have to be continuous, unbounded, and measured on an interval or ratio scale. In the case of TOEFL data, the dependent variable (the essay score) is discrete and bounded between 1 and 6. Because the reported essay score is an average of two raters' ratings, the dependent variable is in increments of 0.5, with 11 valid score categories (i.e., 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0). The second is that it does not require a linear relationship between the dependent and independent variables. Thus, logistic regression allows for the investigation of the group membership effect on the dependent variable, whether the relationships between the dependent and the independent variables are linear or nonlinear. When a dependent variable is discrete and bounded, while the independent variable is continuous, a nonlinear relationship is likely to exist among the variables. In such instances, a logistic regression procedure is the most appropriate method.

The logistic regression method employed in this study was the "proportional odds-ratio model" that is also implemented in the SAS logistic procedure (SAS Institute, 1990). A three-step modeling process based on logistic regression (Zumbo, 1999) was used as a main method of analysis along with a residual-based procedure devised for this study. Polytomous essay scores were dichotomized into 10 binary variables according to the cumulative-logit dichotomization

scheme (see Appendix A for more details). The 10 dichotomized essay variables were simultaneously regressed on examinees' ELA scores, the native language dummy group variable (European = 0; East Asian = 1), and the ability-by-group interaction variable in a step-by-step fashion. Equal slopes were assumed for all of the 10 dichotomized variables from the same prompt. More specifically, the ordinal logistic regression analysis was conducted in the following three steps: In Step 1, the matching variable or the conditioning variable (i.e., ELA scores) was entered into the regression equation for all the dichotomized responses (*i*), as in $g_i(x, D) = \beta_{0i} + \beta_1 x$; In Step 2, the group membership (i.e., European versus East Asian) variable was entered, $g_i(x, D) = \beta_{0i} + \beta_1 x + \beta_2 D_m$; In Step 3, the interaction term (i.e., English language ability-by-group) was finally added, $g_i(x, D) = \beta_{0i} + \beta_1 x + \beta_2 D_m + \beta_3 x D_m$. The three nested models in Steps 1-3 can be fitted to the data and compared in terms of model-data fit (expressed in terms of χ^2 statistics) and R^2 coefficients.

To gauge the amount of the group differences (if any), three different kinds of effect sizes from the logistic regression were used in this study: (a) the residual-based effect size; (b) R^2 combined with *p*-values for the χ^2 test and slope parameters; and (c) the group-specific expected score curves. Before the full three-step modeling process was begun, expected essay scores, residual scores, and the residual-based effect sizes were computed for all the prompts by using only the matching variable (i.e., ELA scores) as an independent variable in the regression model. Expected essay scores for individual test takers' ELA scores were computed from the step-one model, $g_i(x, D) = \beta_{0i} + \beta_1 x$. Residual scores were obtained for individual examinees by subtracting their ELA-predicted essay scores from their raw essay scores, and these residual scores were averaged separately for each language group on each prompt. The residual-based effect sizes were computed by dividing the mean residual score difference between the two groups by the pooled standard deviation of the essay scores for both language groups. The residual-based effect size may be viewed as a measure of the standardized group difference after controlling for the ability difference.

The uniform R^2 effect size is basically an increased portion of R^2 after entering the dummy language group variable into the ability-only regression model (Step 1); the nonuniform effect size, an increased portion of R^2 after adding the interaction term in the Step 2 model. The total effect size is the aggregate of the uniform and nonuniform effects. There is some

controversy about just what constitutes small or negligible, moderate or medium, or large effects. Cohen (1988) considered R^2 effect sizes of 0.02, 0.13, and 0.26 as "small," "medium," and "large" effect sizes, respectively, which can also be linked to the group mean score differences of 0.20, 0.50, and 0.80 in standard deviation units. Roussos and Stout (1996) suggested that R^2 differences of 0.035, 0.035 to 0.070, and greater than .070 be considered as "negligible," "moderate," and "large" effects, which were also adopted by Jodoin and Gierl (2001). Zumbo (1999) has suggested a total R^2 effect size of 0.13 as a minimal effect size for a group effect, provided that the two-degree-of-freedom chi-square (χ^2) test between Steps 1 and 3 has a p-value less than or equal to 0.01. Zumbo's (1999) classification scheme of the R^2 values of 0.13 corresponds to a "medium" R^2 effect size in Cohen's (1988) standard.

The group-specific expected score curves were next obtained based on logistic regression curves for the 10 dichotomized responses and the 11 score characteristic curves for those prompts that were flagged because of significant group effects, as shown in Appendixes A and B. For those prompts with significant ability-by-group interaction effects, the two separate group-specific curves cross at some point. For those prompts with no significant group effect, the curves are essentially identical. This can be regarded as a visual measure of the model-based effect sizes to show vividly the patterns of the uniform and nonuniform effects of the native language backgrounds on the essay scores. The vertical distance between the two lines at each ELA score point can be regarded as the expected essay score difference between the examinees of the same English language ability but from the different language groups.

Results

English Language Ability Versus Observed Essay Scores

Descriptive statistics of the 81 prompts used in the analysis are provided to give a general overview of the score information that was used for the logistic regression analysis for the two comparison groups. Table 1 reports overall means and standard deviations of the raw essay and the ELA scores for both the European and East Asian language groups and standardized mean differences between the two groups, when 81 prompts were analyzed together. Figures 1 and 2 show the patterns of difference in the mean ELA and the mean observed essay score between the two groups across the 81 essay prompts.

Table 1

Means and Standard Deviations of English Language Ability and Observed Essay Scores for European and East Asian Language Groups and Standardized Mean Differences

Variable/native language	Sample size	Mean score	Standard deviation	Standardized mean difference (*d*)
TOEFL essay score				
European group	121,494	4.22	0.95	0.48*
East Asian group	132,941	3.78	0.89	
English Language Ability				
European group	121,494	1.09	2.37	0.84*
East Asian group	132,941	–1.00	2.61	

*$p < 0.01$ two-tailed.

As shown in Table 1 and Figures 1 and 2, the ELA and observed essay scores were higher for the European language group than for the East Asian language group for individual prompts and at the aggregate level. At the aggregate level, the standardized mean difference in the ELA observed, .84, is quite significant ($p < 0.01$) and would be viewed as a "large" effect size (Cohen, 1988). The standardized mean essay score difference observed between the European and East Asian language groups was 0.48, which is also statistically significant ($p < 0.01$) and may be viewed as a "medium" effect size.[1] Even at the individual prompt level, there was a consistently higher mean ELA score observed for the European language group.

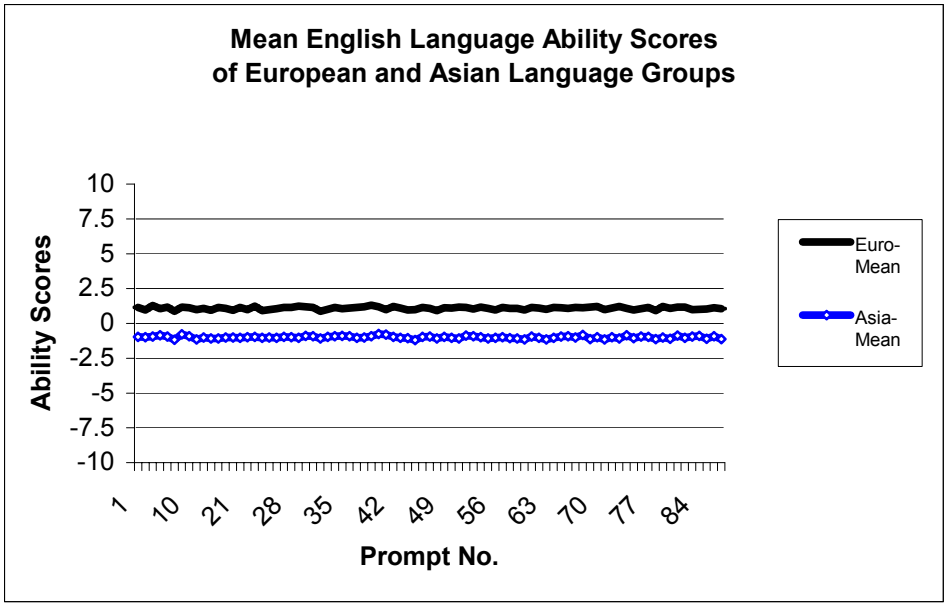

Figure 1. **Mean English language ability of European and East Asian language groups for each of the 81 writing prompts.**

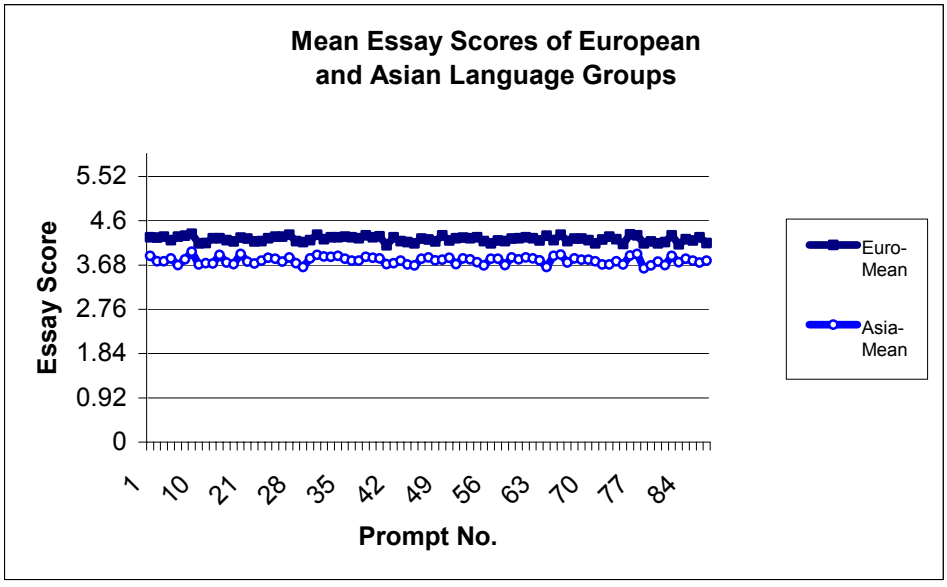

Figure 2. **Mean essay scores of European and East Asian language groups for each of the 81 writing prompts.**

A similar pattern of difference was observed for the mean observed essay scores. It appears that the mean essay score difference between the two groups may be largely attributable to the mean ELA score differences between the two groups. This difference is analogous to what is often called "impact" rather than "item bias" in the DIF literature (Clauser & Mazor, 1998).

A closer inspection of mean essay scores indicates, however, that the distance between the two group score trend lines was less consistent across prompts. Figures 1 and 2 illustrate graphically differences in ELA and essay scores for individual prompts. For a few essay prompts, it seems that the differences in the mean raw essay scores (Figure 2) are not explained by the differences in the mean ELA (Figure 1) between the two groups alone. This suggests that the examinees' native languages may have some observable impact on essay scores.

Residual-based Effect Sizes After Controlling for ELA

Mean expected essay scores from the step-one regression model (only with English ability included as a predictor) were averaged separately for each language group over the 81 prompts and are reported in Table 2. Mean expected scores over the 81 prompts were 4.23 and 3.79 for the European and the East Asian groups, respectively, which were very close to the mean observed essay scores of 4.22 and 3.78. For this reason, the averaged residuals between the observed and expected essays scores across the 81 prompts were also very small (–0.01) both for the East Asian and the European groups, and the averaged effect size turned out to be zero. It may be that the English language ability differences have been already controlled to a large extent by the CBT prompt selection algorithm that assigns a prompt to each examinee in a near-random way.

Table 2

Expected Mean Essay Scores, Residuals, and Standardized Mean Group Differences Averaged Over 81 Prompts After Controlling for English Language Ability Differences Using Logistic Regression (Step 1 Model)

Variable/response mode	Expected score		Residual (observed-expected)		Residual-based effect size (*d*)
	Mean	SD	Mean	SD	
TOEFL Writing Score					
European Group	4.23	0.51	–0.01	0.77	0.00
East Asian Group	3.79	0.55	–0.01	0.71	

Figure 3 illustrates graphically residual scores and effect sizes for individual prompts. The mean residual scores (observed minus expected) for the two groups and the residual-based effect sizes for each of the 81 prompts are visually displayed in Figures 3 and 4, respectively. Figure 3 shows that the European language group performed better than predicted on some prompts (positive residual score) but worse than predicted on others (negative residual score); the same was true for the East Asian language group. In Figure 4, a positive value of the residual-based effect size for a prompt indicates that the prompt is favoring the European group on average, whereas a negative value indicates the other way around. Overall, it seems that the negative and positive effect sizes might be cancelled out across the prompts, if the effect sizes are aggregated across all 81 prompts.

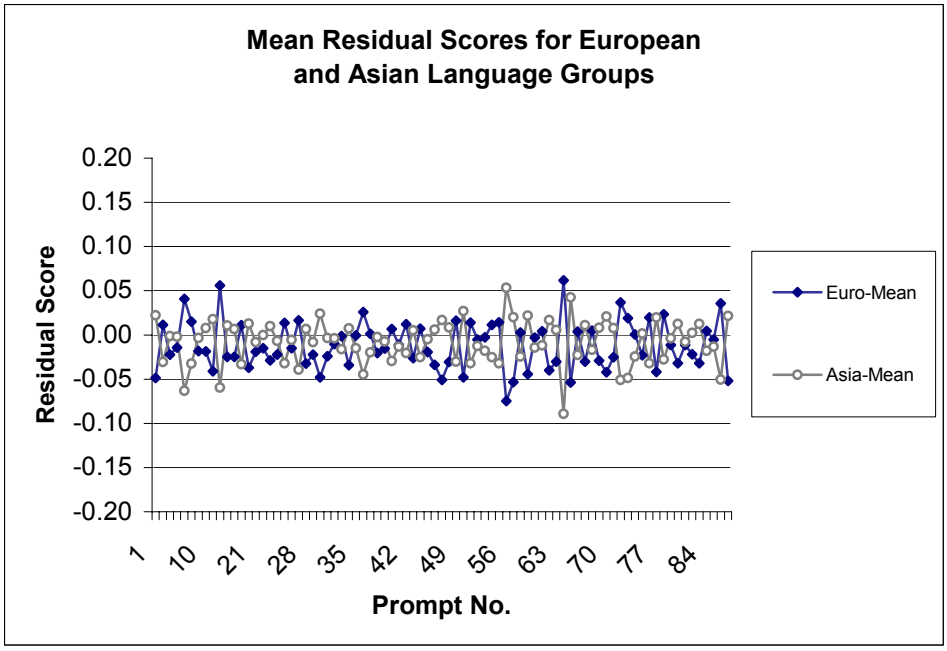

Figure 3. **Mean residual scores after controlling for English language ability of European and East Asian language groups for each prompt.**

A closer examination of each prompt shows, however, that several prompts stand out from other prompts in the magnitude of the effect sizes, as shown in Figure 4. For instance, Prompt 64 had the largest effect size (0.16) and favored the European language group. Prompt 56 had the second largest effect size (about –0.14) and favored the East Asian language group. Both of these two turned out to be less than "small effect sizes" even by Cohen's (1988) 0.20 standard for a "small effect size." These prompts were flagged also in the three-step modeling process that is described in more detail in the next section.

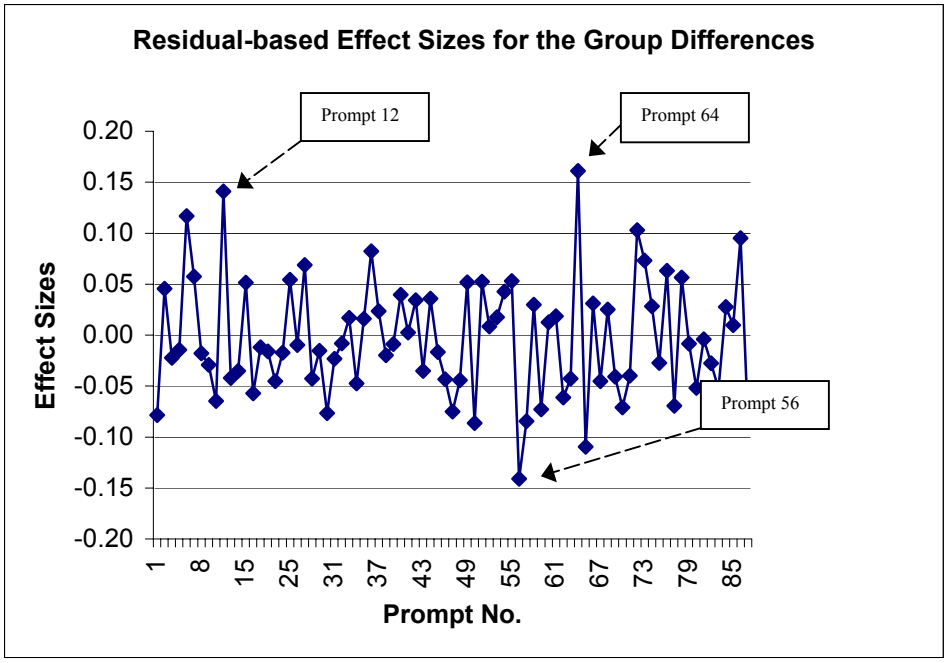

Figure 4. **Residual-based effect sizes of European and East Asian language groups for each prompt.**

Slope Parameters and R^2 Effect Sizes from Three-step Modeling

Table 3 illustrates that the results of the analysis based on the three-step modeling process shows that: (a) the English ability variable (x) was by far the best predictor of essay scores ($p < 0.0001$) for all of the 81 prompts included in this analysis, as expected; (b) the native language effect accounted for a statistically significant amount of variation in essay scores for only 27 of the 81 prompts ($p < 0.05$); and (c) 20 of those 27 prompts exhibited a significant ability-by-group interaction ($x * D_m$) indicating nonuniform effects, but the remaining 7 prompts had a uniform group effect only ($p < 0.05$).

Table 3

Means of Slope Parameters and Increased R^2 Values for Added Predictor Variables in the Logistic Regression for 81 Prompts

Group effect	No. of prompts	English language ability (x)		Native language groups (D_m)		Ability x group interaction (x* D_m)	
		Slope (β_1) mean	Mean R^2	Slope ($\lvert\beta_2\rvert$) mean	Mean R^2	Slope (β_3) mean	Mean R^2
No group effect	54	–0.52**	0.3766				
Uniform only	7	–0.51**	0.3785	0.24*	0.0028		
Uniform-dominant	9	–0.56**	0.3717	0.24*	0.0024	0.06*	0.0012
NU-dominant	11	–0.58**	0.3780	0.17*	0.0011	0.08*	0.0019
Total	81	–0.53**	0.3764				

*$p < 0.05$ two-tailed. **$p < 0.01$ two-tailed.

Uniform R^2 effect sizes for the 27 flagged prompts ranged from 0.0006 to 0.0073, with a mean of 0.0021 and a standard deviation of 0.0017. Shown in Table 4 are the 5 prompts with the largest uniform and nonuniform R^2 effect sizes, respectively. The sizes of the absolute β_2 values were found to be proportional to those of the uniform R^2 effect. However, the direction of the β_2 parameter value was positive for about a half of the prompts (14 prompts), but negative for another half (13 prompts). Prompt 64, for instance, had the largest positive slope value (0.42), favoring the European language to a small degree, whereas Prompt 56 had the largest negative slope value (–0.34), favoring the East Asian language group. All of the 5 prompts in Table 4 turned out to be the same ones that were identified as having the largest residual-based effect

sizes in the previous section. Nevertheless, the sizes of the residual-based effect sizes correspond more closely to the total R^2 effect sizes rather than to the uniform effect size alone.

Table 4

Five Prompts With Largest Uniform R^2 Effect Sizes Estimated From the Three-step Modeling Procedure

Prompt no.	No. of examinees		Slope for language group (β_2)	R^2 effect size		
	European	East Asian		Uniform	Non-uniform	Total
Prompt 64	1,357	1,434	0.42**	0.0073	0	0.0073
Prompt 12	1,087	1,231	0.35**	0.0051	0	0.0051
Prompt 56	1,344	1,576	–0.34**	0.0043	0.0018	0.0061
Prompt 06	1,622	1,737	0.27**	0.0034	0.0015	0.0049
Prompt 65	1,750	1,772	–0.28**	0.0030	0.0011	0.0041

**p < 0.01 two-tailed.

Table 5

Five Prompts With the Largest Nonuniform R^2 Effect Sizes From the Three-step Modeling Procedure

Prompt no.	No. of examinees		Slope for interaction term(β_3)	R^2 effect size		
	European	East Asian		Non-uniform	Uniform	Total
Prompt 67	1,729	1,799	0.11**	0.0032	0.0006	0.0038
Prompt 47	2,335	2,535	0.10**	0.0028	0.0012	0.0040
Prompt 07	1,553	1,722	0.09**	0.0025	0.0010	0.0035
Prompt 57	1,789	2,060	0.09**	0.0022	0.0018	0.0040
Prompt 70	1,620	1,774	0.09**	0.0020	0.0013	0.0033

**p < 0.01 two-tailed.

Nonuniform effects were exhibited in 20 of the 27 prompts with significant uniform group effect, with the nonuniform R^2 effect sizes varying from 0.0006 to 0.0032 among these 20 prompts. When there is a nonuniform effect present in a prompt, it means that the prompt favors

one examinee group over another at low ELA levels, but the opposite is true at high ELA levels (Penfield & Lam, 2000; Swaminathan & Rogers, 1990). However, all of the remaining 7 prompts displaying uniform effect only had positive β_2 values for the D_m variable, consistently favoring the European language group slightly at all ELA levels. Listed in Table 5 are the 5 prompts with the largest nonuniform effects sizes and the largest slope parameters (β_3) for the interaction term ($x^* D_m$). The β_3 values for the interaction term were positive in direction for all of the 20 prompts, and the sizes of the β_3 values were also proportional to the nonuniform R^2 effect sizes. The nonuniform effect was larger than the uniform effect in only 11 of the 20 prompts, but smaller than the uniform effect in the remaining 9 prompts. For the 11 prompts with larger nonuniform effects, the nonuniform R^2 effect size ranged from 0.0010 to 0.0032.

Nevertheless, the increased R^2 values due to the group variable, the interaction variable, or both variables were far too small for any prompts to be regarded as displaying a serious level of uniform, nonuniform, or combined group effects. Among the 5 prompts shown in Table 4, Prompt 64 had the largest uniform and total R^2 effect sizes of 0.0073. On the other hand, Prompt 67 had the largest nonuniform effect size of 0.0032 and the total effect size of 0.0038 among the 5 prompts listed in Table 5. All of these effect sizes are quite small by Cohen's criterion of a small R^2 effect size of 0.02, and they are negligible by the 0.035 standard proposed by Roussos and Stout (1996) or the 0.13 standard proposed by Zumbo (1999).

Group-specific Expected Essay Score Curves

To illustrate the direction and magnitude of the group effects visually, separate expected score curves were drawn for the European and East Asian group for the 27 prompts that had significant uniform group effects. In the case of the 20 prompts with significant group main and ability-group interaction effects together, the slope parameters for both the group (β_2) and the interaction variable (β_3) were entered along with the intercept (β_{0i}) and the slope parameters (β_1) for the ELA variable into the equations for computing (a) logistic regression curves for the dichotomized responses (Figures B1 and B2), (b) category characteristic curves (Figure B3), and (c) group-specific expected essay curves for each group.

The directional and crossover patterns of the group effect on each prompt are more clearly illustrated by drawing separate expected score curves for each language group, as shown in Figures 5 and 6.[2] Figure 5 shows, for example, that the European language group is predicted

to score higher than the East Asian group consistently at all ELA score levels on Prompt 64. The vertical distance between the two score lines was the largest (0.18) at an ELA score point of about –7.4, whereas it was the smallest (0.08) at an ELA score point of 10. When the distances between the two curves were averaged across the actual ELA score range of about –10 to 5 in the data, the average distance was about 0.18, which is close to the residual-based effect size (0.16) previously computed for this prompt.

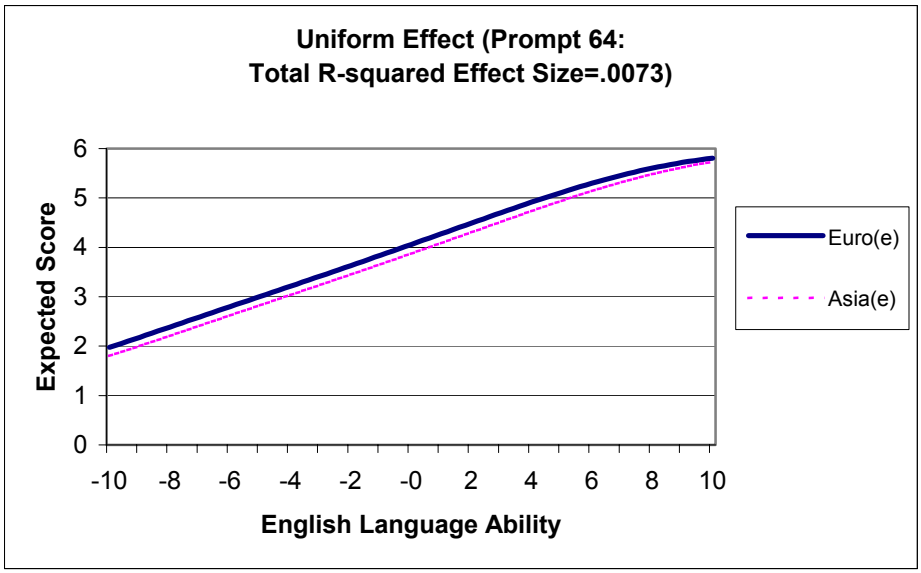

Figure 5. **Uniform effect favoring European language group (Prompt 64) in separate expected score curves for reference and focal groups based on full logistic regression model.**

On the other hand, Figure 6 shows that for Prompt 56, the East Asian language group tended to be advantaged considerably at lower levels of ELA, but not at higher levels. The vertical distance between the two expected score lines was the largest (–0.42) at an ELA score point of about –8.6, whereas the two curves crossed over at the ELA level of about 4.2. When the distances between the two curves were averaged across the actual ELA score range of –10 to 5 in the data, the average distance was about –0.22, which is a bit larger in magnitude than the residual-based effect size (–0.14), but in the same direction, favoring the East Asian language group overall.

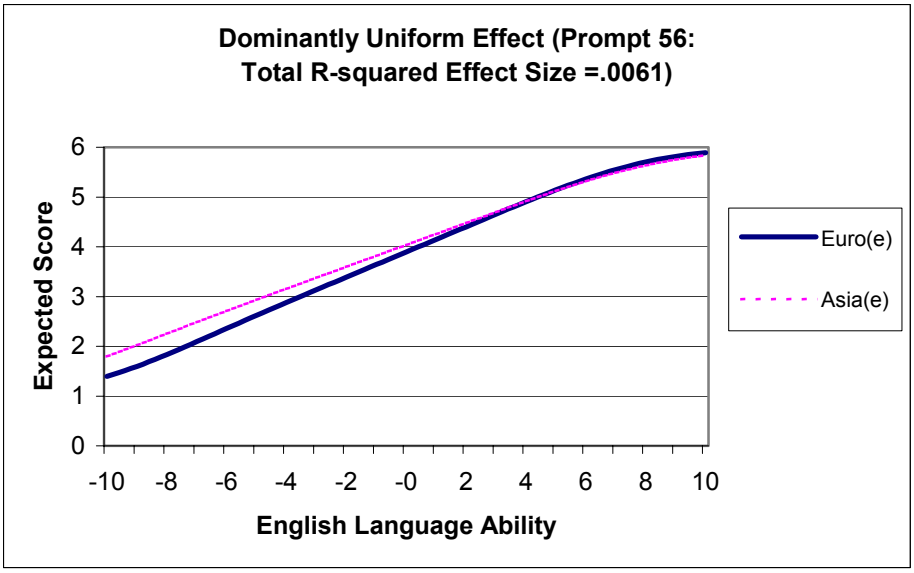

Figure 6. **Dominantly uniform effect favoring Asian Language Group (Prompt 56) in separate expected score curves for reference and focal groups based on full logistic regression model.**

A general pattern of expected scores for the prompts with both uniform and nonuniform effects was that the expected essay scores of the East Asian group were higher at lower ELA levels, but lower than, or equal to, those of the European language group at higher ELA levels. Nevertheless, the magnitude of the expected score difference varied from prompt to prompt, depending upon the direction and size of the slope parameter (β_2) values for the dummy group variable (D_m). Figures 7 and 8 present graphic representations for two prompts with dominantly nonuniform effects. Figure 7 shows a graphic representation for Prompt 67, and Figure 8 provides a graphics representation for Prompt 07. Three prompts in Figures 6, 7, and 8 represent three subtly different patterns of nonuniform effects among the writing prompts: (a) group membership had larger effects at low levels of English language ability (ELA), but group effects disappeared at higher levels of ELA (Prompt 56); (b) group membership had larger effects at low ELA levels, but group effects were reversed at higher levels of ELA (Prompt 67); and (c) group membership had smaller effects at low levels of ELA, and group effects were reversed at higher ELA levels (Prompt 07). For all of these three patterns, examinees of East Asian language group were expected to score higher than those of the European group at low levels of ELA.

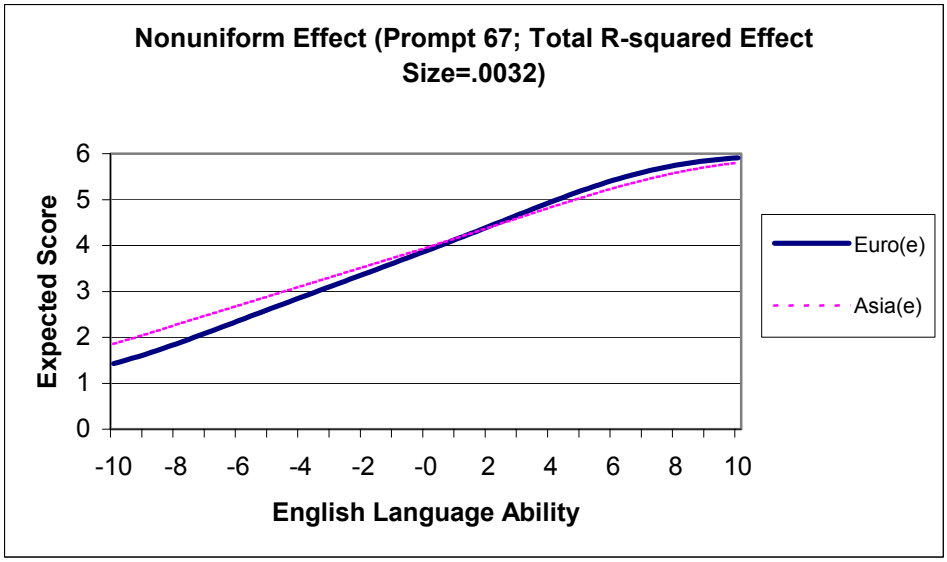

Figure 7. **Dominantly nonuniform effects favoring East Asian language group, especially at lower ability levels (Prompt 67), in separate expected score curves for reference and focal groups based on full logistic regression model.**

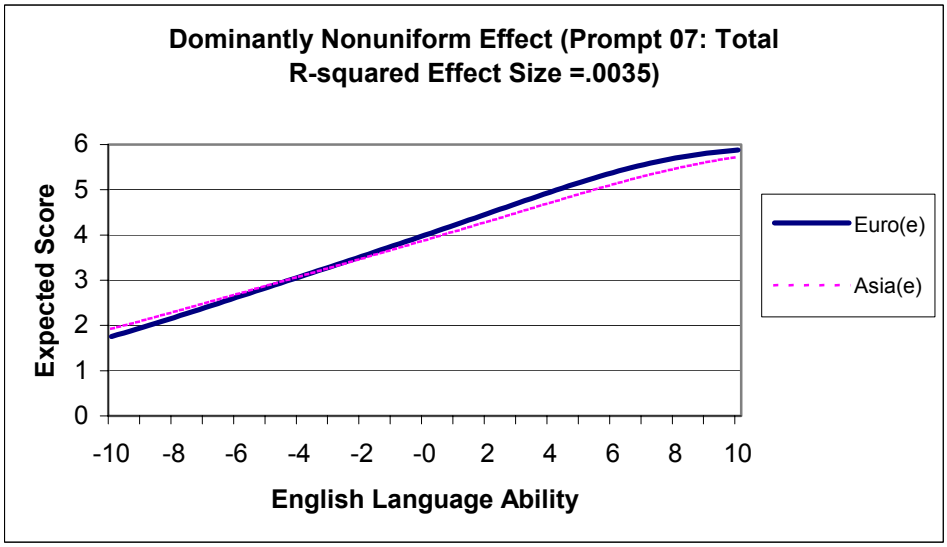

Figure 8. **Dominantly nonuniform effects favoring European language group overall, especially at higher ability levels (Prompt 07), in separate expected score curves for reference and focal groups based on full logistic regression model.**

Summary and Discussion

The number of tasks that can be feasibly administered is usually small in performance-based writing assessments, because such formats of assessment require typically extended responses and are time-consuming to administer (Powers & Fowles, 1998). Often only one prompt is administered to each examinee, as in TOEFL CBT Writing. Under such circumstances, it is very important to ensure that each prompt is as fair as possible to examinee subgroups. The purpose of the present study was to investigate the comparability of TOEFL CBT writing prompts for examinees of European and East Asian native languages. A preliminary examination of the pattern of the mean ELA and raw essay scores revealed that the mean essay score difference between the two groups might be largely ascribable to the difference in mean ELA scores between the two groups. It was also found that, while one third of the prompts were flagged because of statistically significant group differences, the effect sizes were far too small for any of the flagged prompts to be classified as exhibiting an important group effect. The findings of the study are discussed in terms of: (a) item impact and prompt bias; (b) the magnitude of group effects for the flagged items; and (c) the patterns of group effects.

First, essay score differences between the European and East Asian language groups seem to be similar to *item impact* rather than a *group difference* attributable to a construct-irrelevant factor inherent in writing prompts. A clear distinction is usually made between item impact and DIF in the item bias literature (Clauser & Mazor, 1998; Holland & Thayer, 1988; Penfield & Lam, 2000; Zumbo, 1999). Item impact may be present when examinees from different groups have different probabilities of success on an item, because examinees from these groups do differ in ability of interest. In such circumstances, group differences in examinee performance on the item are to be expected because of true differences between the groups in the underlying ability being measured by the item. In this study, a consistently higher mean essay score was observed for the European language group across all of the 81 prompts. Intriguingly enough, the mean ELA score for the European language group was also higher across the 81 prompts. This indicates that the mean essay score difference between the two groups may be largely explained by the mean ELA score difference between the two groups. In other words, examinees of European languages would be expected to score higher on most TOEFL CBT writing prompts largely because they are of higher English language proficiency.

Second, despite such a clear, general pattern in both ELA and observed essays scores, it was found that the mean observed essay score differences between the two groups were less uniform than the ELA score differences. Thus, it was necessary to examine whether the same pattern of difference can be ascertained after examinees of the two groups were matched on ELA. To make a fair comparison of group performance, examinees of the two groups had to be first matched on the relevant underlying ability before determining whether examinees of the two groups differ in their probability of success on the item (Angoff, 1993; Hambleton, Swaminathan, & Rogers, 1991; Zumbo, 1999). It was found that about one third of the prompts were initially flagged due to a statistically significant group effect. However, the R^2 effect sizes of any of the flagged prompts were far too small to be classified as exhibiting an important group effect. For instance, Prompt 64 had the largest total R^2 effect size of 0.0073, which was much smaller than the DIF criterion of 0.13 suggested by Zumbo (1999) and negligible by Cohen's (1988) and Roussos and Stout's (1996) standards as well. Even when the group effect was examined at different ELA levels for those prompts exhibiting nonuniform effects, the largest expected score difference was too small to be indicative of any a serious group effect. For instance, Prompt 56 had the largest nonuniform effect, but the largest vertical distance between the expected score curves in the valid range of ELA scores was smaller than one half of a pooled standard deviation of observed essay scores for the two groups (0.46 standard deviation). This means that all the prompts analyzed in this study were free of any serious group effect even after the examinees of the two groups were matched on ELA, whether or not their performance was examined at different levels of ELA or at the aggregate level in the valid range of ELA scores.

Third, most of the flagged prompts (74%) had nonuniform as well as uniform effects to some extent, whereas all of the remaining prompts displayed only uniform effect favoring the examinees of the European language group consistently at all ELA levels. For each of those prompts displaying nonuniform effect, examinees of the East Asian language group were expected to score higher than those of the European group to a varying degree at the low ELA levels. Nonetheless, a closer inspection has revealed that there are two subtly different patterns of the nonuniform effect: (a) group membership has larger effects at low levels of English language ability (ELA), but group effects disappeared at higher levels of ELA; (b) group membership has smaller or larger effects at low ELA levels, but group effects were reversed at higher levels of ELA. The first pattern may be linked to a hypothesis in second language acquisition (SLA) that the

influence of the first language is generally greatest at the initial stage of SLA (or lower second language proficiency levels), but likely to diminish as second language proficiency increases (Ryan & Bachman, 1992). To put it another way, test performance of test takers whose second language (English) had not been sufficiently developed tend to be influenced more by the test takers' native language backgrounds. This may be also related to the findings of previous studies on the effect of test takers' native language backgrounds on the dimensionality of the TOEFL test across the different levels of language proficiency (Oltman, Stricker, & Barrows, 1988), which suggested that different sections of the TOEFL test would form more differentiated factors for low-proficiency examinees, but less distinct constructs for high-proficiency examinees.

Nonetheless, group effects did not disappear at high ELA levels on all the prompts displaying nonuniform effect. On some prompts, the group effect was reversed instead of being diminished at the higher ELA levels. As pointed out by Ryan and Bachman (1992), the native language group variable might be a surrogate for cultural, social, and educational as well as linguistic differences between the European and East Asian language groups. In relation to this, one important area for further investigation might be a closer examination of impact of test preparation or test-taking strategies on examinees' writing scores for different native language groups. Somewhat contrary to our initial expectation, examinees of the East Asian language group seem to perform better than their counterparts of the European language group at low ELA levels on all prompts exhibiting nonuniform effects. In fact, a large pool of TOEFL CBT prompts are prepublished in the TOEFL Bulletin so that examinees may have an equal chance of becoming familiar with the writing prompts. Then, an intriguing question is what kind of test-taking strategies examinees of different native languages use to compensate for their low writing ability or ELA. It may be possible that some examinees can somehow compensate for their low ELA by using a strategy of memorizing a template of an exemplary essay and replacing some key words in the essay for a new writing prompt.

A related question might be whether the impact of test preparation or test-taking strategies mediates the impact of the native language background differentially depending upon ELA levels on some prompts. One of the notable linguistic differences between the European and the East Asian language groups may be differences in rhetorical conventions (e.g., deductive versus inductive, writer-responsible versus reader responsible, direct vs. indirect) in their first language (L1) writing (Connor, 2002; Hinds, 1990; Hinkel, 1997, 2002; Kincaid, 1987; Scollon

& Scollon, 1991; Taylor, 1995). It would be interesting to investigate whether examinees' L1 rhetorical conventions could transfer to their second language writing, either causing interference or facilitation. This question may be particularly relevant for high-ELA examinees who might be able to write longer essays in which their ability to use appropriate rhetorical structures can be more clearly demonstrated and for which it would be harder to get higher scores simply by using test-taking strategies.

Conclusions, Limitations, and Recommendations for Further Investigation

The findings of this study shows that TOEFL writing prompts as a whole are comparable for examinees of both East Asian and European languages included in this study. Two thirds of the prompts analyzed in this study were found to be free of any statistically significant native-language-related group effects. Even the remaining prompts that were flagged because of statistically significant uniform or nonuniform group effects had effect sizes too small to be indicative of any serious group effects. When the direction of the group effect was examined, the effect was positive for some flagged prompts but negative for others. The total effect size became almost zero at the aggregate level of the total 81 prompts analyzed in this study, because the positive and negative effect sizes were cancelled out. These findings from the logistic regression analyses seem to be consistent with the general pattern of mean ELA and mean essay scores for the 81 prompts. The mean essay score difference between the two groups may be largely attributable to the difference in the mean ELA scores between the two groups. This means that examinees of European languages would be expected to score higher on most TOEFL CBT writing prompts largely because they are of higher English language proficiency.

From a methodological point of view, this study indicated that extension of the logistic regression DIF methodology for use with polytomous items was effective in investigating both uniform and nonuniform group effects related to native languages. It was demonstrated that the directional and crossover patterns as well as the magnitude of group effects for each prompt could be illustrated with separate expected score curves for the two language groups. Such a refinement of the logistic DIF methodology for expected score comparison for the reference and focal groups is one of the unique contributions of the current study. The study may help to dispel concerns raised by some researchers about the utility of logistic regression procedures for polytomous items (French & Miller, 1996; Kim, 2001).

Nonetheless, further investigations in the following three areas would prove very valuable in deepening our understanding of the impact of native languages on performance on TOEFL CBT writing prompts. First, one limitation of the present study was that the ELA variable used in this study may not be an ideal measure of writing ability with which to match examinees of different native language groups. Because no internal matching criterion for writing ability was available in the same writing section, the ELA variable had to be used as a matching variable representing examinees' writing proficiency, on the assumption that, if examinees had high English language ability measured by the three sections of the test as a whole, they would be also expected to perform well overall on the essay scores, and vice versa. However, a remaining question may be whether the logistic regression procedure using ELA was more (or less) sensitive to a potential group effect than a more direct writing measure would have been.

Second, even when significant *statistical* group effects were detected, that condition alone is not sufficient to indicate prompt bias. A conclusion of item (or prompt) bias "requires the further condition that the observed difference in item performance can be attributed to some property of the item that is unrelated to the construct intended to be measured by the test" (Penfield & Lam, 2000). Prompt content analysis was not pursued in this study because the R^2 effect size even for the most extreme prompt was found to be too small to warrant further analysis. Nonetheless, the practical significance of the effect sizes may also be judgmental by nature and dependent on the test purposes (Kim, 2001; Prentice & Miller, 1992). If the test being investigated is high-stakes with respect to the impact of the decision on test takers, then the results of DIF analyses should be examined more carefully with respect to the content of the prompt.

In addition to the small effect sizes for the flagged writing prompts, there was another important reason for not pursuing the prompt content analysis in this study. In this study, the examinees of three European languages (French, German, and Spanish) formed a reference group, while those of three East Asian languages (Chinese, Japanese, and Korean) formed a focal group. It should be noted, however, that the definition of the reference and focal groups based on examinees' native languages are rather fuzzy and not as clear-cut as those based on gender (male/female) or response modes (typed/handwritten) in the context of TOEFL. The *TOEFL Bulletin* actually lists more than 100 native languages. Even though the classification scheme

used in this study may be justifiable on several valid grounds (e.g., Indo-European vs. non-Indo-European language family, Roman alphabet-based vs. non-Roman-alphabet-based writing system, large sample sizes for both focal reference groups), this may still be arbitrary to some extent. Even within each of the two language groups, subtle differences exist among individual languages. Although we may find some construct-irrelevant feature of the prompt causing a significant group effect for these two language groups, it may not provide a conclusive piece of evidence for revision or exclusion of certain prompts from the pool, for there may be some other large language groups (e.g., Arabic, Thai) that were not accounted for in the analysis.

Third, raters can also be a source of the native-language-related group effect as well as the content characteristics of prompts (Henning, 1996; Lynch & McNamara, 1998; Miller & Linn, 2000). In our study, the averaged ratings over two raters were used as a dependent variable. It may be important, however, to disentangle the prompt-related and rater-related effects in the examination of the group effect associated with native language backgrounds. In this sense, the rater behavior and the rating process might need to be carefully studied along with the content of the prompts through substantive analysis (Kim, 2001) in the context of TOEFL CBT writing.

With all these considered, it is recommended that a policy should be clearly formulated for what might be important comparison groups in the context of TOEFL CBT and what levels of difference should result in prompts being dropped from active administration. Even though a one-time expert review of prompts can indicate why some prompts may be less comparable than others, it is a relatively inefficient procedure and it does not always explain why differences occur. Thus, it is also recommended that statistical quality control procedures based on the logistic regression method used in this study be routinely implemented to identify less comparable prompts for various predefined focal groups in terms of examinee native language. Prompt developers can benefit from routinely identifying prompts through statistical quality control and then reviewing those that are identified as extreme. From a methodological perspective, an exploration may also need to be made to find an appropriate methodology that can examine multiple focal groups simultaneously (Penfield, 2001).

References

Angoff, W. H. (1989). *Context bias in the Test of English as a Foreign Language* (TOEFL Research Report No. 29). Princeton, NJ: ETS.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum.

American Educational Research Association, American, Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Breland, H., Muraki, E., & Lee, Y. (2001, April). *Comparability of TOEFL CBT writing prompts for different response modes.* Paper presented at the annual meeting of National Council on Measurement in Education (NCME) in Seattle, WA.

Breland, H., Muraki, E., Lee, Y., Najarian, M., & Beyer, J. (2000, April). *Comparability in TOEFL CBT essay prompts using the logistic regression method.* Paper presented at the annual conference of National Council on Measurement in Education (NCME) in New Orleans, LA.

Brown, A., & Iwashita, N. (1998). The role of language background in the validation of a computer-adaptive test. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 195-207). Mahwah, NJ: Lawrence Erlbaum.

Chang, H., Mazzeo, J., & Roussos, L. A. (1995). *Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure* (ETS RR-95-5). Princeton, NJ: ETS.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, *2*, 155-163.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Connor, U. (2002). New directions in contrastive rhetoric. *TESOL Quarterly, 36*(4), 493-510.

DeMauro, G. (1992). *An investigation of the appropriateness of the TOEFL test as matching variable to equate TWE topics* (TOEFL Research Report No. 37). Princeton, NJ: ETS.

Dorans, N. J., & Schmitt, A. J. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS RR-91-47). Princeton, NJ: ETS.

ETS. (1998). *Computer-based TOEFL score user guide*. Princeton, NJ: Author.

ETS. (1999). *TOEFL tips: Preparing students for the computer-based test*. Princeton, NJ: Author.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*, 315-332.

Gentile, C., Riazantseva, A., & Cline, F. (2001). *A comparison of handwritten and word-processed TOEFL essays.* Manuscript submitted for publication.

Ginther, A., & Stevens, J. (1998). Language background, ethnicity, and the internal construct validity of the Advanced Placement Spanish Language Examination. In Kunnan, A. J. (Ed.), *Validation in language assessment* (pp. 169-94). Mahwah, NJ: Lawrence Erlbaum.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). *Confirmatory factor analysis of the Test of English as a Foreign Language* (TOEFL Research Report No. 32). Princeton, NJ: ETS.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hauck, W. W. (1983). A note on confidence bands for the logistic response curve. *The American Statistician*, *37*, 158-160.

Henning, G. (1996). Accounting for nonsystematic error in performance testing. *Language Testing*, *13*, 53-61.

Hinds. J. (1990). Inductive, deductive, quasi-inductive: Expository writing in Japanese, Korean, Chinese, and Thai. In U. Connor & A. Johns (Eds.), *Coherence in writing* (pp. 87-110). Alexandria, VA: TESOL.

Hinkel, E. (1997). Indirectness in L1 and L2 academic writing. *Journal of Pragmatics*, *27*(*3*), 360-386.

Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features*. Mahwah, NJ: Lawrence Erlbaum.

Holland, P., & Thayer, D. (1988). Differential item performance. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*, 89-114.

Kincaid, L. (Ed.). (1987). *Communication theory: Eastern and Western perspectives*. San Diego, CA: Academic Press.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-80.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement*, *24,* 367-378.

Miller, T., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, *30*, 107-122.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple group partial credit model. *Journal of Educational Measurement*, *36*(3), 217-232.

Oltman, P. K., Stricker, L. J., & Barrows, T. (1988). *Native language, English proficiency, and the structure of the Test of English as a Foreign Language for several language groups*. (TOEFL Research Report No. 27). Princeton, NJ: ETS.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenzel Procedures. *Applied Measurement in Education*, *14*(3), 235-259.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, *19*(3), 5-15.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, *19*, 23-37.

Powers, D. E., & Fowles, M. E. (1998). *Test takers' judgments about GRE writing test prompts* (ETS RR-98-36). Princeton, NJ: ETS.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*(*1*), 160-164.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenzel Type I error performance. *Journal of Educational Measurement, 33,* 215-230.

Ryan, K. E., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, *9*(1), 12-29.

Samejima, F. (1997). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph, 17.*

SAS Institute. (1990). *SAS/STAT user's guide, version six* (4th ed., Vol. 2). Cary, NC: Author.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement. *Language Testing, 8*(2), 95-111.

Scollon, R., & Scollon, S. W. (1991). Topic confusion in English-Asian discourse. *World Englishes*, *10*(2), 113-125.

Sheppard, L. A. (1982). Definition of bias. In R. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9-30). Baltimore, MD: John Hopkins University.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Swinton, S. S., & Powers, D. E. (1980). *Factor analysis of the Test of English as a Foreign Language for several language groups* (TOEFL Research Report No. 6). Princeton, NJ: ETS.

Taylor, I. (1995). *Writing and literacy in Chinese, Korean, and Japanese.* Amsterdam: John Benjamins.

Wainer, H., Serici, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*, 197-219.

Wolfe, E. W., & Manalo, J. R. (2001). *An investigation of the impact of composition medium on the quality of scores from the TOEFL writing section: A report from the broad-based study.* Manuscript submitted for publication.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type*

(ordinal) item scores. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233-251.

Notes

[1] One of the reviewers has pointed out that the group difference for the essay scores (0.48) is likely smaller than the group difference for the ELA scores (0.84) largely because of difference in reliability. We may agree that the reliability difference could have partially contributed to the difference in magnitude of effect sizes based on the essay and ELA scores and that this result should be interpreted with that caution in mind. Nevertheless, it would be fairer to say that the reliability difference between the matching variable and the studied item is inherent in almost all of the DIF studies, because more reliable total scores (or some function of them) are usually used as a matching variable. It may be the case that scores from a performance-based test may be less reliable than those from the traditional MC standardized test in general, because a much smaller number of items can be given in each form of the test (sometimes a single item) and because subjective human judgment is involved in the scoring. Nonetheless, it is our belief that a single writing prompt scored by two independent raters according to a six-point rubric is not necessarily less reliable than a single MC item scored dichotomously.

[2] From a technical point of view, it would be possible to draw 95% confidence intervals around the expected essay score curves for the focal and reference groups based on a procedure outlined by Hauck (1983) and implemented by Miller and Spray (1993) in the *logistic discriminant function* analysis. As one reviewer has pointed out, the confidence intervals might have indicated that the two different expected curves for the East Asian and European groups might have been indistinguishable even for those prompts with the largest group effects. However, we should point out that it would require a more complex, multi-step procedure to draw confidence intervals around the expected essay score curves (not logistic curves for dichotomized responses) in the logistic regression than logistic discriminant function analyses (e.g., creating confidence intervals for (a) logistic regression curves for 10 dichotomized responses; (b) score characteristic curves for 11 score categories; and (c) expected score curves). Moreover, variance and covariance components among the variables have to be estimated from the logistic regression. Even though we recognize it as an important area for further technical refinement for the logistic regression procedure employed in this study, we do not believe that this would change the general conclusion of the study that none of the prompts analyzed would be classified as exhibiting a serious group effect in terms of the effect sizes.

Appendix A

Derivation of the Logistic Regression Model for Polytomous Items: The Proportional Odds-ratio Model

The multiple logistic regression equations for dichotomous items (i) can be written as:

$$P(U_i \mid x, D) = \frac{\exp[g_i(x, D)]}{1 + \exp[g_i(x, D)]} = \frac{1}{1 + \exp[-(g_i(x, D))]} \quad (1)$$

where U_i represents the binary responses for dichotomized items i (U_i = 0 or 1) and x is the continuous variable score, and D is the design matrix of the covariate variables. In this equation, the function g_i (x,D) is called a *logit*. The logit is a linear combination of the continuous score (x), a covariate variable (D), and an interaction term (xD). If we want to analyze the DIF for M levels of a native language covariate, as in our TOEFL essay data, we can rewrite the logit g_i (x,D) as:

$$g_i(x, D) = \beta_{0i} + \beta_1 x + \beta_2 D_m + \beta_3 x D_m \quad (2)$$

where β_{0i} is the intercept for a dichotomous item (i), β_1 is the slope parameter associated with the English language ability score, β_2 is the parameter associated with the native language group variable, D_m, and β_3 is the slope parameter associated with the ability score-by-group interaction. In our study, D_m is 0 for the European language group and 1 for the East Asian language group, respectively. It should be noted that the score-by-group interaction term was added to examine the score difference of nonuniform nature between the two groups.

The dichtomous model in Equation 1 can be directly extended for a polytomous item case based on the cumulative logit dichotomization scheme (Agresti, 1990; French & Miller, 1996). For the polytomous case, K+1 response categories for the polytomous item are dichotomized into K binary responses, and then the logistic regression is fitted to each dichotomized response for the ordinal item, with the parallel slopes assumed for all the dichotomized responses. In the actual TOEFL CBT essay data, there are 11 valid reported score categories (e.g., 1, 1.5, …, 5.5, 6), and, thus, there are 10 dichotomized responses. The proportional log-odds for each dichotomized response based on the cumulative logit scheme can be expressed as:

$$L_{ijk} = \ln[\frac{\Pr(y_j \le k \mid x,D)}{1-\Pr(y_j \le k \mid x,D)}] = \ln[\frac{P_0(x,D)+P_1(x,D)+.....+P_k(x,D)}{P_{k+1}(x,D)+P_{k+2}(x,D)+.....+P_K(x,D)}] \quad (3)$$

where L_{jk} stands for the proportional log-odds ratio for a dichotomized response (i) on the polytomous item (j), k is a subscript of the response category (k = 0,1,2, …, K) for examinee scores (y) on the polytomous essay item, j. It should be noted that in this scheme the proportional log-odds ratio for this dichotomized response for Prompt j is $\Pr(y_j \le k \mid x, D)$ over $[1-\Pr(y_j \le k+1 \mid x, D)]$, which is the opposite of Samejima's (1997) graded response model.

Category Characteristic Curves

If we define $P^+_{jk}(x,D)$ and $P^+_{j,k+1}(x,D)$ as the regression of the binary item score method in which all score categories smaller than k and k+1, respectively, are scored 0 for each dichotomized item, the actual score category characteristic curve for score category k of the graded item j in relation to the independent variables x is

$$P_{jk}(x,D) = P^+_{j,k+1}(x,D) - P^+_{jk}(x,D) \quad (4)$$

where

$$P^+_{jk}(x,D) = \sum_{v=0}^{k} P_{jv}(x,D)$$

Since the differencing scheme based on the cumulative logit logistic regression should be the opposite of Samejima's (1997) scheme, $P^+_{j0}(x,D)$ and $P^+_{j,K+1}(x,D)$ can be also defined in such a way that

$$P^+_{j0}(x,D) = 0$$

and

$$P^+_{j,K+1}(x,D) = 1$$

In the TOEFL CBT essay data, the score category response model for $y_j = k$ can be expressed by

$$P_{jk}(x, D) = \frac{\exp[(g_{j,i+1}(x, D)]}{1 + \exp[(g_{j,i+1}(x, D)]} - \frac{\exp[(g_{ji}(x, D)]}{1 + \exp[(g_{ji}(x, D)]} \tag{5}$$

Appendix B

Logistic Regression Curves for Dichotomized Responses

Figure B1 through B3 show the logistic regression curves for dichotomized responses and category characteristic curves derived from the odds-ratio model.

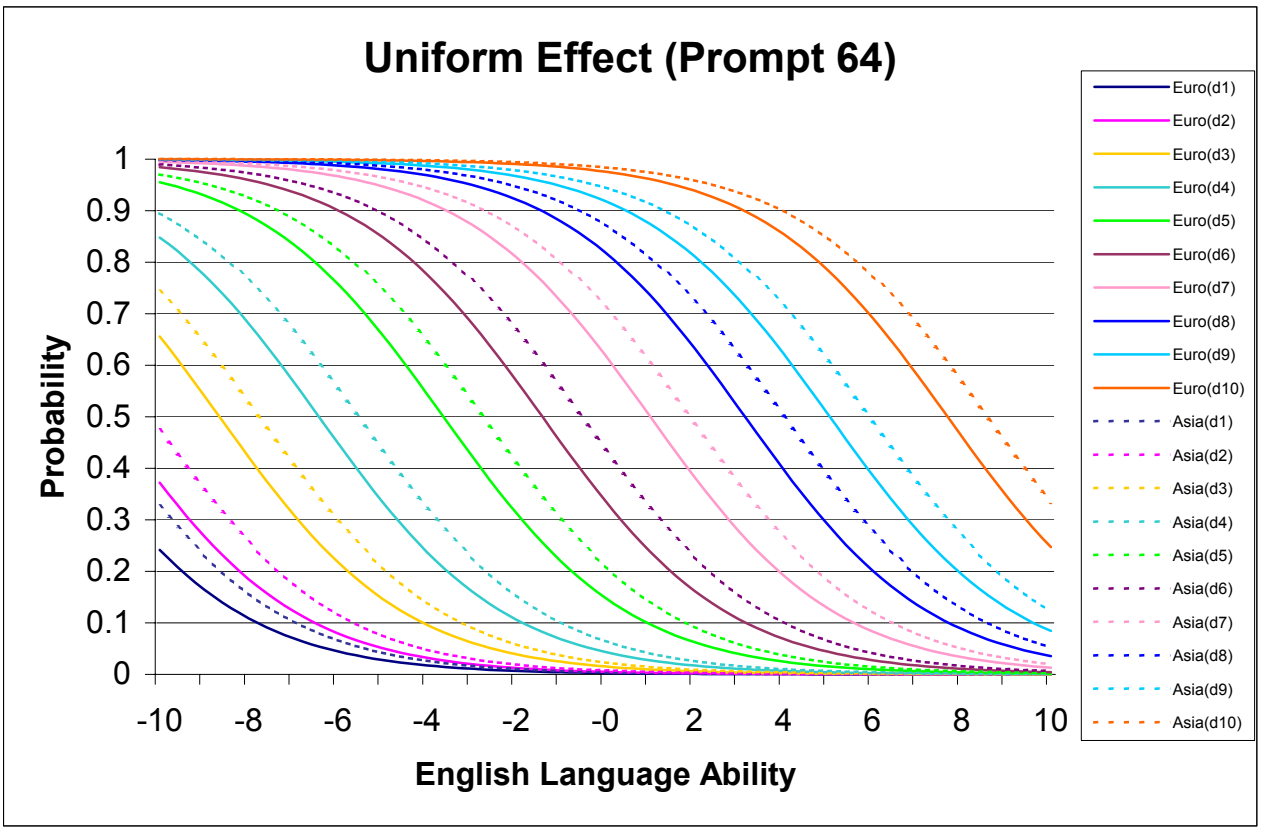

Figure B1. **Logistic regression curves for 10 dichotomized item responses (0) for European and East Asian language groups on Prompt 64.**

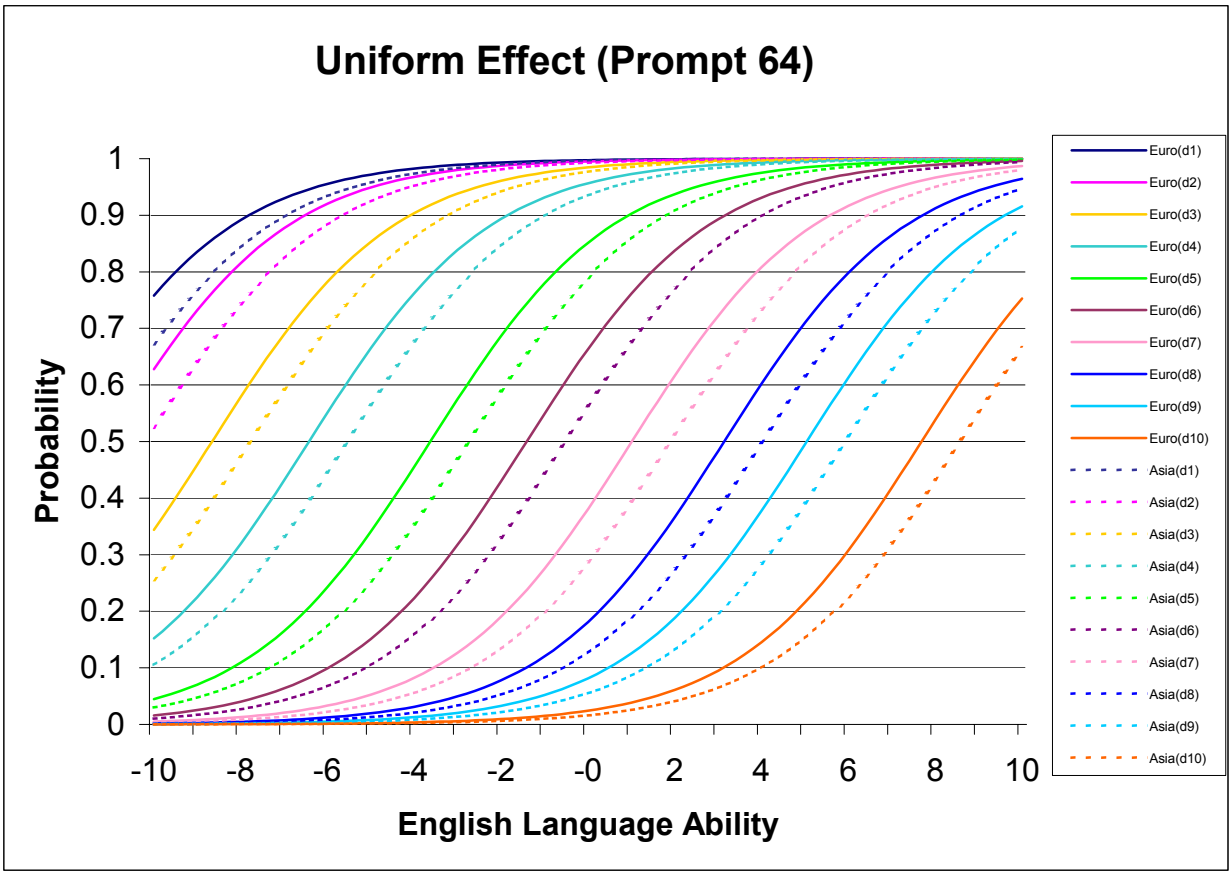

Figure B2. **Logistic regression curves for 10 dichotomized item responses (1) for European and East Asian language groups on Prompt 64.**

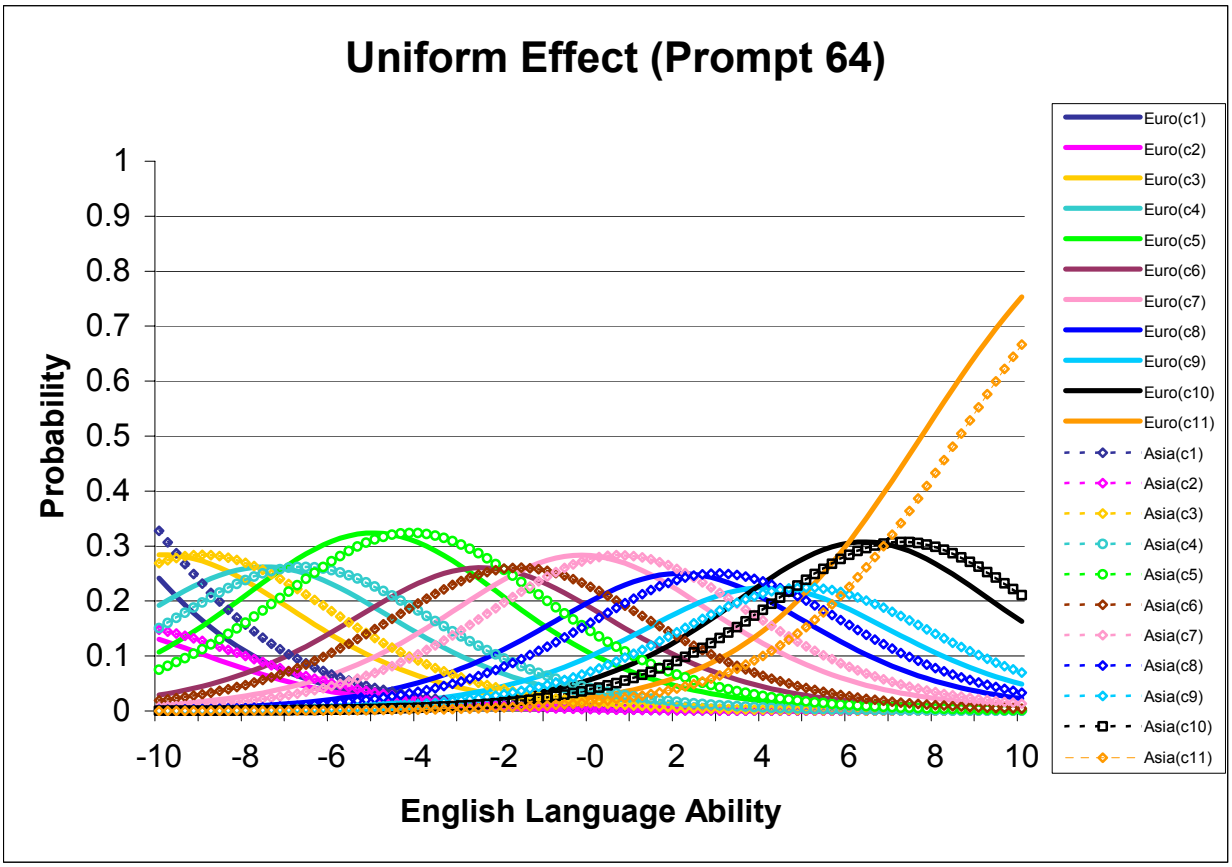

Figure B3. **Score category characteristic curves for 11 score categories for European and East Asian language groups on Prompt 64.**

Appendix C

Number of Essays, Means and Standard Deviations of Observed Essay Score

Tables C1 through C3 show the number of essays, means, and standard deviations of observed essay scores. The tables also show the means of standard deviations of the English language ability (ELA) scores. Six native language subgroups for 81 prompts are represented in the tables.

Table C1

Number of Examinees of Six Native Language Subgroups for 81 Prompts

Prompt	European Language Group				East Asian Language Group				Total
no.	French	German	Spanish	Subtotal	Chinese	Japanese	Korean	Subtotal	*N*
1	270	219	614	1,103	468	414	383	1,265	2,368
3	362	359	924	1,645	705	631	505	1,841	3,486
4	380	375	867	1,622	668	564	483	1,715	3,337
5	470	427	1,035	1,932	817	731	569	2,117	4,049
6	359	378	885	1,622	682	544	511	1,737	3,359
7	404	319	830	1,553	637	610	475	1,722	3,275
8	226	196	506	928	425	322	282	1,029	1,957
10	350	334	782	1,466	621	440	430	1,491	2,957
11	359	338	870	1,567	697	604	509	1,810	3,377
12	220	232	635	1,087	498	396	337	1,231	2,318
13	267	247	677	1,191	519	419	350	1,288	2,479
14	218	248	676	1,142	515	450	407	1,372	2,514
15	482	481	1,100	2,063	833	788	573	2,194	4,257
19	268	235	645	1,148	521	410	441	1,372	2,520
21	347	359	812	1,518	643	548	440	1,631	3,149
22	220	271	689	1,180	538	431	402	1,371	2,551
23	410	400	890	1,700	680	539	483	1,702	3,402
24	373	337	826	1,536	624	580	446	1,650	3,186
25	350	327	801	1,478	626	564	408	1,598	3,076
26	252	238	611	1,101	492	403	331	1,226	2,327
27	388	377	936	1,701	684	544	486	1,714	3,415
28	263	239	593	1,095	461	376	331	1,168	2,263
29	198	250	517	965	411	361	262	1,034	1,999
30	377	356	779	1,512	614	505	424	1,543	3,055
31	324	310	757	1,391	615	508	456	1,579	2,970

(Table continues)

Table C1 (continued)

Prompt	European Language Group				East Asian Language Group				Total
no.	French	German	Spanish	Subtotal	Chinese	Japanese	Korean	Subtotal	*N*
32	286	275	676	1,237	500	479	376	1,355	2,592
33	460	462	1,103	2,025	821	711	584	2,116	4,141
34	421	403	962	1,786	742	602	532	1,876	3,662
35	337	371	842	1,550	645	585	419	1,649	3,199
36	285	243	613	1,141	503	402	367	1,272	2,413
37	327	317	743	1,387	559	453	383	1,395	2,782
38	440	431	1,052	1,923	805	725	608	2,138	4,061
39	307	310	744	1,361	551	424	383	1,358	2,719
40	364	338	802	1,504	583	438	451	1,472	2,976
41	444	407	964	1,815	710	593	544	1,847	3,662
42	385	379	877	1,641	699	576	519	1,794	3,435
43	340	335	824	1,499	687	545	478	1,710	3,209
44	402	350	1,003	1,755	831	673	614	2,118	3,873
45	338	334	783	1,455	642	586	448	1,676	3,131
46	410	443	988	1,841	782	648	545	1,975	3,816
47	511	494	1,330	2,335	1,046	816	673	2,535	4,870
48	267	237	593	1,097	465	414	334	1,213	2,310
49	504	478	1,165	2,147	961	821	708	2,490	4,637
50	470	481	1,118	2,069	856	740	606	2,202	4,271
51	341	303	842	1,486	716	537	518	1,771	3,257
52	374	383	867	1,624	668	618	496	1,782	3,406
53	348	292	670	1,310	562	427	401	1,390	2,700
54	275	277	699	1,251	597	494	382	1,473	2,724
55	298	308	813	1,419	644	549	447	1,640	3,059
56	294	301	749	1,344	641	511	424	1,576	2,920
57	378	402	1,009	1,789	817	678	565	2,060	3,849
58	390	398	966	1,754	739	608	478	1,825	3,579
59	481	482	1,131	2,094	869	809	602	2,280	4,374
60	309	266	652	1,227	543	471	402	1,416	2,643
61	404	428	993	1,825	768	610	518	1,896	3,721
62	251	239	583	1,073	495	387	356	1,238	2,311
63	386	320	846	1,552	665	611	511	1,787	3,339
64	325	323	709	1,357	559	481	394	1,434	2,791
65	379	401	970	1,750	680	612	480	1,772	3,522
66	398	390	964	1,752	760	603	529	1,892	3,644
67	391	408	930	1,729	745	570	484	1,799	3,528
68	339	315	747	1,401	606	499	512	1,617	3,018
69	374	362	883	1,619	695	567	460	1,722	3,341

(Table continues)

Table C1 (continued)

Prompt	European Language Group				East Asian Language Group				Total
no.	French	German	Spanish	Subtotal	Chinese	Japanese	Korean	Subtotal	*N*
70	356	360	904	1,620	710	576	488	1,774	3,394
71	399	325	828	1,552	669	567	518	1,754	3,306
72	519	503	1,212	2,234	902	736	690	2,328	4,562
73	317	359	801	1,477	664	556	461	1,681	3,158
74	332	351	804	1,487	620	510	420	1,550	3,037
75	363	387	915	1,665	747	567	514	1,828	3,493
76	265	253	678	1,196	558	446	438	1,442	2,638
77	366	363	874	1,603	682	569	492	1,743	3,346
78	311	264	649	1,224	541	485	424	1,450	2,674
79	305	313	718	1,336	498	480	370	1,348	2,684
80	274	215	538	1,027	490	390	288	1,168	2,195
81	320	331	779	1,430	612	466	388	1,466	2,896
82	361	366	842	1,569	649	499	469	1,617	3,186
83	209	216	537	962	436	350	325	1,111	2,073
84	211	212	516	939	396	307	286	989	1,928
85	236	199	593	1,028	504	448	384	1,336	2,364
86	220	213	536	969	421	352	358	1,131	2,100
87	473	437	1,096	2,006	862	807	595	2,264	4,270
Total	28,007	27,205	66,282	121,494	52,112	43,666	37,163	132,941	254,435
M	346	336	818	1,500	643	539	459	1,641	3,141
SD	77	78	178	329	134	121	94	343	667

Table C2

Mean ELA Scores of Six Language Subgroups for 81 Prompts

	European Languages								Asian Languages							
Prompt	French		German		Spanish		Euro-Tot		Chinese		Japanese		Korean		Asia-Tot	
no.	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
1	0.98	2.19	1.93	1.94	0.94	2.37	1.15	2.28	–0.39	2.52	–1.73	2.74	–0.87	2.39	–0.97	2.62
3	0.43	2.57	2.21	1.85	0.70	2.48	0.97	2.47	–0.57	2.62	–1.41	2.71	–1.06	2.41	–0.99	2.62
4	1.17	2.20	2.30	1.72	0.90	2.48	1.28	2.33	–0.41	2.63	–1.56	2.77	–0.96	2.52	–0.94	2.69
5	0.75	2.35	2.10	1.78	0.78	2.41	1.06	2.34	–0.40	2.59	–1.34	2.73	–0.89	2.43	–0.86	2.63
6	0.91	2.52	2.17	1.84	0.86	2.49	1.18	2.42	–0.56	2.50	–1.40	2.80	–1.00	2.47	–0.95	2.61
7	0.74	2.50	1.85	2.06	0.55	2.51	0.87	2.47	–0.78	2.55	–1.68	2.52	–1.02	2.26	–1.17	2.49
8	1.17	2.29	2.10	1.76	0.82	2.45	1.18	2.33	–0.40	2.49	–1.34	2.67	–0.76	2.57	–0.79	2.59
10	0.90	2.23	2.19	1.77	0.75	2.48	1.12	2.35	–0.46	2.51	–1.66	2.46	–0.85	2.42	–0.93	2.52
11	0.60	2.47	2.18	1.68	0.69	2.53	0.99	2.44	–0.76	2.58	–1.73	2.60	–1.03	2.47	–1.16	2.59

(Table continues)

Table C2 (continued)

Prompt	European Languages								Asian Languages							
	French		German		Spanish		Euro-Tot		Chinese		Japanese		Korean		Asia-Tot	
no.	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
12	0.59	2.64	2.27	1.65	0.84	2.27	1.09	2.32	–0.54	2.57	–1.52	2.56	–1.14	2.33	–1.02	2.54
13	0.84	2.28	2.26	1.82	0.50	2.53	0.94	2.44	–0.58	2.63	–1.74	2.64	–1.02	2.53	–1.08	2.65
14	0.87	2.29	2.15	1.88	0.86	2.49	1.14	2.39	–0.55	2.51	–1.77	2.61	–0.96	2.41	–1.07	2.56
15	0.74	2.31	2.05	1.75	0.79	2.45	1.07	2.33	–0.55	2.57	–1.70	2.71	–0.77	2.38	–1.02	2.63
19	0.63	2.43	2.26	1.78	0.61	2.54	0.95	2.47	–0.73	2.79	–1.53	2.63	–0.89	2.53	–1.02	2.68
21	0.96	2.22	2.15	1.80	0.79	2.27	1.15	2.22	–0.54	2.62	–1.69	2.79	–0.96	2.51	–1.04	2.70
22	0.81	2.70	2.16	1.87	0.57	2.64	0.98	2.58	–0.37	2.65	–1.62	2.58	–1.11	2.53	–0.98	2.65
23	0.81	2.46	2.36	1.60	0.91	2.38	1.23	2.33	–0.52	2.55	–1.65	2.80	–0.83	2.32	–0.97	2.61
24	0.50	2.53	1.85	1.86	0.72	2.51	0.91	2.44	–0.52	2.57	–1.62	2.68	–0.98	2.53	–1.03	2.64
25	0.58	2.34	2.20	1.94	0.69	2.41	1.00	2.39	–0.56	2.61	–1.48	2.72	–1.11	2.40	–1.02	2.63
26	0.80	2.29	2.28	1.67	0.69	2.49	1.06	2.37	–0.62	2.62	–1.72	2.54	–0.82	2.35	–1.03	2.57
27	0.76	2.28	2.23	1.70	0.88	2.47	1.15	2.35	–0.64	2.67	–1.45	2.58	–0.86	2.38	–0.96	2.58
28	0.88	2.27	2.13	1.96	0.85	2.34	1.14	2.30	–0.46	2.56	–1.63	2.62	–0.99	2.38	–0.98	2.58
29	0.82	2.42	2.24	1.76	0.89	2.45	1.23	2.36	–0.55	2.72	–1.75	2.76	–0.84	2.36	–1.04	2.70
30	0.79	2.52	2.18	1.78	0.94	2.39	1.20	2.36	–0.47	2.57	–1.52	2.68	–0.78	2.57	–0.90	2.64
31	1.00	2.34	2.31	1.67	0.72	2.40	1.14	2.33	–0.62	2.57	–1.41	2.49	–0.79	2.34	–0.93	2.50
32	0.72	2.43	2.06	1.87	0.47	2.51	0.88	2.44	–0.47	2.51	–1.66	2.45	–1.20	2.44	–1.09	2.52
33	0.79	2.48	2.07	1.94	0.67	2.50	1.02	2.45	–0.59	2.59	–1.48	2.76	–0.89	2.38	–0.97	2.62
34	1.00	2.35	2.08	1.90	0.81	2.29	1.14	2.28	–0.48	2.56	–1.47	2.65	–0.90	2.45	–0.92	2.59
35	0.82	2.47	2.19	1.77	0.65	2.44	1.06	2.39	–0.45	2.49	–1.40	2.63	–0.91	2.26	–0.90	2.52
36	0.79	2.55	2.17	1.75	0.83	2.38	1.11	2.37	–0.42	2.48	–1.50	2.59	–1.00	2.47	–0.93	2.55
37	0.71	2.50	2.18	1.72	0.91	2.44	1.15	2.38	–0.70	2.71	–1.46	2.65	–1.03	2.42	–1.04	2.63
38	0.85	2.37	2.30	1.75	0.88	2.36	1.19	2.32	–0.66	2.64	–1.54	2.63	–0.82	2.56	–1.01	2.64
39	0.94	2.32	2.32	1.76	1.06	2.33	1.32	2.27	–0.48	2.62	–1.63	2.66	–0.72	2.38	–0.91	2.61
40	0.71	2.43	2.31	1.60	0.95	2.33	1.20	2.29	–0.27	2.60	–1.43	2.74	–0.75	2.54	–0.76	2.67
41	0.70	2.52	2.17	1.86	0.65	2.44	1.00	2.43	–0.32	2.47	–1.39	2.62	–0.82	2.39	–0.81	2.53
42	1.02	2.56	2.25	1.69	0.86	2.28	1.21	2.30	–0.45	2.61	–1.53	2.76	–1.02	2.42	–0.96	2.65
43	0.72	2.34	2.16	1.73	0.83	2.41	1.10	2.33	–0.55	2.47	–1.66	2.70	–1.02	2.38	–1.03	2.56
44	0.69	2.39	2.03	1.91	0.69	2.49	0.96	2.42	–0.63	2.58	–1.61	2.69	–1.05	2.56	–1.06	2.64
45	0.60	2.36	2.07	1.89	0.70	2.33	0.99	2.32	–0.79	2.51	–1.63	2.76	–1.19	2.47	–1.19	2.61
46	0.81	2.51	2.19	1.70	0.84	2.29	1.16	2.29	–0.52	2.64	–1.57	2.73	–0.92	2.38	–0.97	2.64
47	0.87	2.56	2.33	1.71	0.68	2.41	1.07	2.41	–0.46	2.61	–1.55	2.77	–0.94	2.39	–0.94	2.65
48	0.69	2.39	2.08	1.84	0.57	2.51	0.93	2.43	–0.68	2.58	–1.65	2.69	–0.98	2.63	–1.09	2.66
49	0.90	2.34	2.14	1.67	0.79	2.45	1.12	2.34	–0.44	2.59	–1.63	2.60	–0.93	2.43	–0.97	2.60
50	0.79	2.38	2.31	1.79	0.71	2.54	1.10	2.44	–0.61	2.63	–1.58	2.74	–0.95	2.37	–1.03	2.63
51	1.00	2.32	2.23	1.91	0.88	2.40	1.18	2.35	–0.51	2.58	–1.79	2.54	–1.08	2.37	–1.07	2.56
52	0.93	2.40	2.32	1.76	0.75	2.39	1.16	2.35	–0.43	2.50	–1.44	2.70	–0.78	2.51	–0.88	2.61

(Table continues)

Table C2 (continued)

	European Languages								East Asian Languages							
Prompt	French		German		Spanish		Euro-Tot		Chinese		Japanese		Korean		Asia-Tot	
no.	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
53	0.86	2.32	2.10	1.78	0.67	2.47	1.04	2.36	–0.61	2.60	–1.40	2.61	–0.81	2.60	–0.91	2.62
54	0.82	2.28	2.24	1.80	0.88	2.36	1.17	2.30	–0.59	2.45	–1.51	2.61	–0.95	2.21	–0.99	2.48
55	0.70	2.63	2.12	1.72	0.83	2.38	1.08	2.37	–0.69	2.55	–1.62	2.51	–0.95	2.49	–1.07	2.55
56	0.69	2.32	2.18	1.85	0.57	2.52	0.96	2.43	–0.76	2.54	–1.60	2.73	–0.80	2.46	–1.04	2.61
57	0.79	2.37	2.27	1.79	0.84	2.42	1.15	2.36	–0.48	2.56	–1.66	2.74	–0.88	2.51	–0.98	2.65
58	0.88	2.49	1.98	1.92	0.77	2.43	1.07	2.39	–0.70	2.67	–1.49	2.60	–1.05	2.51	–1.05	2.63
59	0.76	2.52	2.02	1.84	0.81	2.37	1.08	2.35	–0.53	2.61	–1.70	2.66	–1.01	2.49	–1.07	2.64
60	0.82	2.25	1.92	1.85	0.65	2.46	0.97	2.34	–0.88	2.64	–1.56	2.68	–0.99	2.46	–1.14	2.62
61	1.08	2.28	2.25	1.80	0.70	2.49	1.15	2.38	–0.43	2.62	–1.60	2.63	–0.97	2.57	–0.95	2.65
62	0.65	2.52	2.28	1.69	0.83	2.40	1.11	2.38	–0.50	2.55	–1.76	2.79	–0.98	2.42	–1.03	2.64
63	0.91	2.21	2.17	1.84	0.61	2.57	1.01	2.42	–0.73	2.61	–1.69	2.70	–1.10	2.49	–1.16	2.64
64	0.95	2.53	2.15	1.89	0.81	2.42	1.16	2.40	–0.46	2.57	–1.71	2.71	–1.00	2.30	–1.03	2.60
65	0.96	2.30	2.28	1.61	0.72	2.53	1.13	2.39	–0.53	2.54	–1.55	2.71	–0.78	2.42	–0.95	2.60
66	0.75	2.34	2.01	1.91	0.84	2.33	1.08	2.30	–0.51	2.62	–1.57	2.70	–0.77	2.52	–0.92	2.66
67	1.02	2.19	2.09	1.91	0.77	2.60	1.14	2.42	–0.57	2.58	–1.60	2.67	–1.04	2.35	–1.02	2.58
68	0.86	2.39	2.02	1.80	0.87	2.33	1.13	2.29	–0.26	2.48	–1.47	2.59	–0.84	2.43	–0.82	2.55
69	0.85	2.31	2.01	1.83	0.95	2.40	1.17	2.31	–0.79	2.55	–1.60	2.66	–1.03	2.58	–1.12	2.62
70	0.84	2.30	2.20	1.80	0.96	2.31	1.21	2.27	–0.53	2.56	–1.67	2.50	–0.85	2.53	–0.99	2.58
71	0.62	2.55	2.15	1.74	0.72	2.37	0.99	2.38	–0.88	2.68	–1.56	2.77	–1.06	2.28	–1.15	2.61
72	0.94	2.42	2.22	1.65	0.71	2.50	1.10	2.39	–0.46	2.66	–1.67	2.72	–0.94	2.42	–0.99	2.66
73	1.02	2.44	2.42	1.73	0.76	2.37	1.22	2.35	–0.64	2.59	–1.66	2.77	–0.98	2.50	–1.07	2.66
74	0.93	2.29	2.26	1.75	0.65	2.40	1.09	2.33	–0.45	2.52	–1.51	2.68	–0.69	2.42	–0.86	2.59
75	0.65	2.36	2.01	1.88	0.65	2.46	0.96	2.39	–0.61	2.60	–1.76	2.65	–0.91	2.36	–1.05	2.59
76	0.77	2.56	2.29	1.78	0.70	2.42	1.05	2.42	–0.41	2.61	–1.63	2.59	–0.94	2.58	–0.95	2.64
77	0.93	2.42	2.37	1.82	0.74	2.50	1.15	2.44	–0.45	2.47	–1.71	2.66	–0.78	2.39	–0.96	2.57
78	0.72	2.29	1.97	1.73	0.59	2.55	0.92	2.40	–0.63	2.60	–1.52	2.67	–1.30	2.36	–1.12	2.59
79	0.94	2.29	2.05	1.96	0.97	2.39	1.22	2.32	–0.52	2.67	–1.65	2.75	–0.88	2.47	–1.02	2.69
80	0.58	2.65	2.09	1.84	0.91	2.26	1.07	2.36	–0.87	2.64	–1.45	2.60	–1.02	2.42	–1.10	2.58
81	0.91	2.26	2.22	1.67	0.83	2.40	1.17	2.29	–0.41	2.63	–1.44	2.66	–0.96	2.39	–0.88	2.61
82	0.95	2.27	2.29	1.83	0.78	2.35	1.17	2.30	–0.65	2.65	–1.50	2.67	–1.04	2.55	–1.03	2.65
83	0.54	2.58	2.05	1.77	0.71	2.53	0.98	2.46	–0.43	2.61	–1.76	2.61	–0.73	2.34	–0.94	2.60
84	0.58	2.60	2.38	1.70	0.64	2.47	1.02	2.46	–0.42	2.50	–1.59	2.57	–0.79	2.56	–0.89	2.59
85	0.83	2.36	2.06	2.01	0.76	2.51	1.03	2.44	–0.50	2.57	–1.89	2.64	–0.97	2.34	–1.10	2.60
86	0.93	2.32	2.26	1.70	0.76	2.33	1.13	2.28	–0.50	2.61	–1.60	2.56	–0.79	2.59	–0.93	2.63
87	0.81	2.26	2.25	1.76	0.68	2.39	1.05	2.32	–0.68	2.53	–1.74	2.69	–0.92	2.46	–1.12	2.61
Total	0.81	2.40	2.17	1.79	0.77	2.43	1.09	2.37	–0.55	2.58	–1.59	2.66	–0.93	2.44	–1.00	2.61

Table C3

Means Observed Essay Scores of Six Language Subgroups for 81 Prompts

	European Language Group								East Asian Language Group							
Prompt	French		German		Spanish		Euro-Tot		Chinese		Japanese		Korean		Asia-Tot	
no.	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
1	4.19	0.91	4.63	0.91	4.15	0.96	4.26	0.96	4.12	0.85	3.66	0.92	3.78	0.89	3.87	0.91
3	4.20	0.96	4.64	0.90	4.12	1.03	4.25	1.01	3.96	0.84	3.61	0.95	3.66	0.96	3.76	0.92
4	4.26	0.84	4.67	0.85	4.11	0.98	4.27	0.95	3.98	0.89	3.57	0.95	3.69	0.93	3.76	0.94
5	4.08	0.89	4.61	0.78	4.10	0.93	4.20	0.91	4.01	0.84	3.68	0.86	3.74	0.85	3.82	0.86
6	4.26	0.92	4.63	0.85	4.11	1.00	4.27	0.97	3.85	0.84	3.58	0.94	3.56	0.87	3.68	0.89
7	4.30	0.93	4.67	0.87	4.14	0.96	4.29	0.96	3.96	0.77	3.73	0.86	3.69	0.81	3.80	0.82
8	4.33	0.81	4.72	0.79	4.17	0.92	4.33	0.89	4.14	0.78	3.84	0.88	3.84	0.86	3.96	0.85
10	4.19	0.88	4.51	0.88	3.94	1.01	4.13	0.98	3.88	0.89	3.47	0.89	3.65	0.91	3.69	0.91
11	4.12	0.88	4.49	0.83	4.02	1.01	4.14	0.96	3.90	0.85	3.58	0.95	3.65	0.90	3.72	0.91
12	4.20	0.83	4.63	0.81	4.12	0.90	4.24	0.89	3.89	0.80	3.56	0.82	3.61	0.81	3.71	0.82
13	4.30	0.80	4.65	0.76	4.08	0.89	4.24	0.87	4.05	0.76	3.73	0.87	3.84	0.88	3.89	0.84
14	4.19	0.87	4.52	0.90	4.08	0.95	4.20	0.94	3.92	0.84	3.55	0.92	3.69	0.90	3.73	0.90
15	4.15	0.85	4.47	0.83	4.05	0.97	4.17	0.93	3.89	0.79	3.51	0.88	3.69	0.87	3.70	0.86
19	4.25	0.88	4.71	0.81	4.09	0.94	4.25	0.93	4.09	0.86	3.79	0.83	3.82	0.91	3.91	0.88
21	4.24	0.88	4.61	0.87	4.05	0.96	4.23	0.95	3.97	0.86	3.56	0.92	3.66	0.85	3.75	0.90
22	4.20	0.93	4.56	0.91	4.01	1.03	4.17	1.01	3.96	0.87	3.57	0.91	3.52	0.93	3.71	0.92
23	4.24	0.87	4.46	0.87	4.02	0.98	4.18	0.94	3.91	0.84	3.64	0.89	3.71	0.82	3.77	0.86
24	4.19	0.90	4.58	0.86	4.12	0.96	4.24	0.94	4.06	0.84	3.63	0.87	3.78	0.93	3.83	0.90
25	4.18	0.80	4.68	0.80	4.14	0.95	4.27	0.91	4.01	0.80	3.67	0.85	3.69	0.85	3.81	0.84
26	4.23	0.85	4.68	0.94	4.12	1.00	4.27	0.98	3.94	0.93	3.53	0.92	3.74	0.97	3.75	0.95
27	4.30	0.82	4.81	0.87	4.11	0.92	4.31	0.93	3.96	0.81	3.78	0.76	3.74	0.83	3.84	0.81
28	4.01	1.00	4.64	0.82	4.07	0.97	4.18	0.97	3.95	0.87	3.56	0.92	3.59	0.90	3.72	0.92
29	4.03	1.00	4.57	0.90	4.01	1.08	4.16	1.05	3.82	0.88	3.46	0.90	3.59	0.93	3.64	0.92
30	4.13	0.88	4.51	0.84	4.09	1.00	4.20	0.95	4.02	0.87	3.59	0.91	3.79	0.99	3.82	0.94
31	4.41	0.92	4.67	0.89	4.12	0.96	4.31	0.96	4.04	0.83	3.73	0.89	3.85	0.91	3.89	0.88
32	4.26	0.84	4.50	0.79	4.09	0.93	4.22	0.89	4.06	0.79	3.75	0.77	3.74	0.81	3.86	0.80
33	4.31	0.90	4.61	0.88	4.10	0.94	4.26	0.94	4.08	0.81	3.68	0.87	3.75	0.82	3.85	0.85
34	4.26	0.90	4.68	0.87	4.08	0.97	4.26	0.96	4.05	0.85	3.69	0.86	3.80	0.89	3.87	0.88
35	4.24	0.88	4.70	0.81	4.10	0.92	4.27	0.92	4.02	0.90	3.63	0.90	3.75	0.87	3.81	0.91
36	4.19	0.88	4.64	0.82	4.14	0.92	4.26	0.91	3.98	0.82	3.60	0.89	3.67	0.80	3.77	0.85
37	4.28	0.88	4.55	0.80	4.10	0.95	4.24	0.92	3.99	0.82	3.60	0.89	3.64	0.90	3.77	0.88
38	4.28	0.86	4.67	0.81	4.16	0.97	4.30	0.93	3.98	0.87	3.72	0.93	3.83	0.93	3.85	0.91
39	4.23	0.81	4.63	0.81	4.11	0.91	4.26	0.89	3.99	0.83	3.68	0.87	3.75	0.89	3.83	0.87
40	4.30	0.96	4.63	0.83	4.14	0.96	4.28	0.95	4.01	0.88	3.63	0.94	3.77	0.87	3.82	0.91
41	4.05	0.93	4.53	0.90	3.93	1.00	4.09	0.99	3.81	0.82	3.61	0.89	3.64	0.88	3.70	0.87
42	4.25	0.87	4.62	0.80	4.10	0.94	4.26	0.92	3.96	0.89	3.53	0.97	3.59	0.94	3.72	0.95

(Table continues)

Table C3 (continued)

Prompt	European Language Group								East Asian Language Group							
	French		German		Spanish		Euro-Tot		Chinese		Japanese		Korean		Asia-Tot	
no.	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
43	4.19	0.79	4.52	0.86	4.04	0.98	4.18	0.93	3.94	0.82	3.62	0.93	3.71	0.90	3.77	0.89
44	4.20	0.92	4.52	0.87	4.01	1.00	4.16	0.98	3.92	0.81	3.51	0.93	3.58	0.93	3.69	0.90
45	4.04	0.94	4.46	0.89	4.02	0.98	4.13	0.97	3.79	0.86	3.58	0.88	3.61	0.87	3.67	0.87
46	4.27	0.90	4.58	0.89	4.06	0.99	4.23	0.97	3.99	0.88	3.66	0.93	3.72	0.92	3.81	0.92
47	4.22	0.92	4.73	0.90	4.01	1.01	4.21	1.01	4.03	0.85	3.69	0.93	3.73	0.89	3.84	0.90
48	4.16	0.90	4.58	0.86	4.01	1.02	4.17	0.98	3.96	0.81	3.66	0.88	3.70	0.93	3.78	0.88
49	4.38	0.93	4.63	0.79	4.13	0.99	4.30	0.95	4.00	0.86	3.60	0.86	3.71	0.92	3.79	0.90
50	4.21	0.88	4.55	0.85	4.03	0.98	4.19	0.95	4.02	0.83	3.67	0.92	3.75	0.81	3.83	0.87
51	4.25	0.93	4.62	0.93	4.10	0.99	4.24	0.98	3.88	0.83	3.48	0.89	3.66	0.87	3.70	0.88
52	4.22	0.88	4.64	0.85	4.09	0.91	4.25	0.92	3.95	0.84	3.70	0.90	3.78	0.89	3.82	0.88
53	4.27	0.93	4.52	0.85	4.10	1.02	4.24	0.97	3.96	0.83	3.61	0.87	3.79	0.89	3.80	0.87
54	4.11	0.92	4.66	0.82	4.17	0.98	4.26	0.96	3.93	0.79	3.59	0.90	3.64	0.85	3.74	0.86
55	4.17	0.98	4.43	0.84	4.09	0.96	4.18	0.95	3.86	0.84	3.56	0.87	3.54	0.88	3.67	0.88
56	4.16	0.86	4.58	0.86	3.93	1.02	4.13	0.99	3.99	0.84	3.64	0.92	3.72	0.95	3.81	0.91
57	4.16	0.88	4.58	0.89	4.04	0.95	4.19	0.95	4.00	0.85	3.67	0.84	3.72	0.89	3.81	0.87
58	4.22	0.93	4.52	0.88	4.02	0.98	4.18	0.97	3.87	0.86	3.51	0.91	3.61	0.93	3.68	0.91
59	4.23	0.94	4.60	0.83	4.07	1.00	4.23	0.97	4.05	0.86	3.68	0.92	3.76	0.89	3.84	0.90
60	4.21	0.89	4.62	0.81	4.09	0.94	4.24	0.92	3.94	0.88	3.67	0.91	3.74	0.88	3.80	0.90
61	4.27	0.86	4.55	0.79	4.12	0.91	4.26	0.89	3.99	0.82	3.69	0.82	3.79	0.88	3.84	0.85
62	4.21	0.87	4.58	0.84	4.11	0.98	4.24	0.94	4.04	0.86	3.68	0.97	3.67	0.92	3.82	0.93
63	4.24	0.89	4.52	0.86	4.05	0.99	4.19	0.96	3.98	0.79	3.64	0.89	3.70	0.82	3.78	0.85
64	4.25	0.95	4.60	0.97	4.17	0.96	4.29	0.98	3.80	0.90	3.42	0.96	3.69	0.90	3.64	0.94
65	4.23	0.88	4.61	0.82	4.01	0.97	4.20	0.95	4.02	0.83	3.78	0.91	3.78	0.88	3.87	0.88
66	4.26	0.84	4.64	0.78	4.20	0.92	4.31	0.89	4.02	0.82	3.73	0.88	3.91	0.81	3.90	0.85
67	4.21	0.94	4.54	0.89	4.02	1.06	4.18	1.02	3.91	0.89	3.53	0.94	3.67	0.84	3.73	0.91
68	4.19	0.83	4.55	0.81	4.11	0.92	4.23	0.89	3.97	0.84	3.71	0.85	3.75	0.85	3.82	0.85
69	4.16	0.88	4.55	0.84	4.13	0.94	4.23	0.92	3.96	0.85	3.68	0.93	3.66	0.95	3.79	0.91
70	4.14	0.89	4.56	0.87	4.09	0.96	4.20	0.94	3.97	0.83	3.59	0.86	3.78	0.93	3.79	0.89
71	4.01	0.90	4.48	0.80	4.04	0.85	4.13	0.87	3.88	0.80	3.65	0.87	3.72	0.78	3.76	0.82
72	4.19	0.89	4.55	0.82	4.08	0.94	4.21	0.92	3.84	0.79	3.53	0.89	3.67	0.85	3.69	0.85
73	4.22	0.97	4.65	0.90	4.11	0.98	4.27	0.98	3.87	0.91	3.52	0.93	3.63	0.90	3.69	0.92
74	4.19	0.85	4.66	0.86	4.05	0.99	4.22	0.96	3.92	0.83	3.59	0.93	3.74	0.85	3.76	0.88
75	4.17	0.88	4.52	0.86	3.93	1.01	4.12	0.98	3.89	0.87	3.51	0.92	3.61	0.85	3.69	0.90
76	4.21	0.77	4.73	0.79	4.20	0.89	4.32	0.87	3.96	0.80	3.83	0.83	3.83	0.83	3.88	0.82
77	4.30	0.88	4.74	0.79	4.13	1.00	4.30	0.96	4.15	0.81	3.72	0.89	3.81	0.90	3.91	0.88
78	4.20	0.85	4.48	0.81	3.96	1.00	4.13	0.95	3.80	0.84	3.49	0.92	3.51	0.91	3.61	0.90
79	4.16	0.89	4.53	0.91	4.02	0.96	4.17	0.95	3.88	0.88	3.46	0.94	3.65	0.93	3.67	0.93

(Table continues)

Table C3 (continued)

Prompt	European Language Group								East Asian Language Group							
	French		German		Spanish		Euro-Tot		Chinese		Japanese		Korean		Asia-Tot	
no.	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
80	4.10	0.88	4.38	0.84	4.05	0.92	4.13	0.90	3.92	0.80	3.64	0.87	3.63	0.88	3.75	0.85
81	4.18	0.89	4.59	0.82	3.97	1.01	4.16	0.97	3.90	0.91	3.52	0.94	3.53	0.91	3.68	0.94
82	4.27	0.89	4.67	0.84	4.16	0.92	4.30	0.92	4.04	0.82	3.77	0.90	3.75	0.89	3.87	0.88
83	4.06	0.83	4.42	0.85	4.01	1.01	4.11	0.96	3.97	0.79	3.53	0.85	3.67	0.84	3.74	0.84
84	4.15	0.95	4.70	0.79	4.06	0.94	4.22	0.94	3.96	0.76	3.68	0.80	3.73	0.84	3.81	0.81
85	4.26	0.81	4.51	0.78	4.06	0.95	4.19	0.91	3.97	0.82	3.52	0.84	3.78	0.80	3.77	0.84
86	4.29	0.91	4.56	0.79	4.13	0.85	4.26	0.87	3.90	0.85	3.57	0.89	3.69	0.95	3.73	0.90
87	4.14	0.87	4.60	0.85	3.95	0.96	4.14	0.95	3.96	0.82	3.59	0.88	3.72	0.87	3.77	0.87
Total	4.21	0.89	4.59	0.85	4.08	0.97	4.22	0.95	3.96	0.84	3.62	0.89	3.70	0.88	3.78	0.88

Appendix D

Mean Expected Essay Scores, Residuals, and Standardized Mean Group Differences

Table D1 shows the mean expected essay scores, residuals, and standardized mean group differences after controlling for English language ability differences using the Step 1 logistic regression model:

$$\left(P(U_j \mid x, D) = \frac{1}{1 + \exp[-(\beta_{0i} + \beta_1 x)]} \right)$$

Table D1

Mean Expected Essay Scores and Residual-based Effect Sizes

	Expected essay scores				Residual (observed-expected)				Mean	Pooled	Residual
Prompt	European		East Asian		European		East Asian		resid.	SD	effect
no.	M	SD	M	SD	M	SD	M	SD	diff.	(obs)	size
1	4.31	0.50	3.84	0.57	0.05	0.78	0.02	0.72	–0.07	0.91	–0.08
3	4.24	0.56	3.79	0.60	0.01	0.79	–0.03	0.75	0.04	0.92	0.05
4	4.29	0.56	3.76	0.64	–0.02	0.76	0.00	0.69	–0.02	0.94	–0.02
5	4.22	0.48	3.83	0.53	–0.01	0.74	0.00	0.70	–0.01	0.86	–0.01
6	4.23	0.55	3.74	0.58	0.04	0.79	–0.06	0.72	0.10	0.89	0.12
7	4.27	0.53	3.84	0.53	0.01	0.77	–0.03	0.66	0.05	0.82	0.06
8	4.35	0.45	3.97	0.50	–0.02	0.75	0.00	0.69	–0.02	0.85	–0.02
10	4.15	0.53	3.68	0.57	–0.02	0.78	0.01	0.74	–0.03	0.91	–0.03
11	4.19	0.54	3.71	0.58	–0.04	0.76	0.02	0.72	–0.06	0.91	–0.06
12	4.19	0.46	3.77	0.50	0.06	0.75	–0.06	0.68	0.12	0.82	0.14
13	4.27	0.48	3.88	0.52	–0.02	0.69	0.01	0.68	–0.04	0.84	–0.04
14	4.22	0.54	3.72	0.58	–0.02	0.76	0.01	0.69	–0.03	0.90	–0.04
15	4.16	0.47	3.74	0.53	0.01	0.78	–0.03	0.70	0.04	0.86	0.05
19	4.29	0.49	3.90	0.52	–0.04	0.76	0.01	0.72	–0.05	0.88	–0.06
21	4.24	0.50	3.76	0.59	–0.02	0.77	–0.01	0.70	–0.01	0.90	–0.01
22	4.19	0.62	3.71	0.63	–0.01	0.76	0.00	0.69	–0.01	0.92	–0.02
23	4.21	0.47	3.76	0.52	–0.03	0.78	0.01	0.69	–0.04	0.86	–0.05
24	4.26	0.53	3.84	0.57	–0.02	0.76	–0.01	0.71	–0.02	0.90	–0.02
25	4.26	0.49	3.84	0.53	0.01	0.75	–0.03	0.68	0.05	0.84	0.05
26	4.28	0.60	3.76	0.64	–0.01	0.76	–0.01	0.72	–0.01	0.95	–0.01
27	4.29	0.46	3.88	0.50	0.02	0.79	–0.04	0.67	0.06	0.81	0.07
28	4.21	0.54	3.71	0.60	–0.03	0.78	0.01	0.71	–0.04	0.92	–0.04
29	4.18	0.56	3.64	0.63	–0.02	0.81	–0.01	0.72	–0.01	0.92	–0.02

(Table continues)

Table D1 (continued)

Prmpt no.	Expected essay scores: European M	European SD	East Asian M	East Asian SD	Residual (observed-expected): European M	European SD	East Asian M	East Asian SD	Mean resid. diff.	Pooled SD (obs)	Residual effect size
30	4.25	0.51	3.79	0.57	–0.05	0.76	0.02	0.75	–0.07	0.94	–0.08
31	4.33	0.50	3.89	0.53	–0.02	0.80	0.00	0.72	–0.02	0.88	–0.02
32	4.23	0.46	3.86	0.47	–0.01	0.74	0.00	0.67	–0.01	0.80	–0.01
33	4.26	0.49	3.87	0.52	0.00	0.77	–0.02	0.71	0.01	0.85	0.02
34	4.30	0.49	3.86	0.54	–0.03	0.79	0.01	0.71	–0.04	0.88	–0.05
35	4.27	0.54	3.83	0.57	0.00	0.75	–0.02	0.70	0.01	0.91	0.02
36	4.23	0.49	3.81	0.52	0.03	0.75	–0.04	0.71	0.07	0.85	0.08
37	4.24	0.50	3.79	0.55	0.00	0.75	–0.02	0.72	0.02	0.88	0.02
38	4.32	0.49	3.85	0.56	–0.02	0.77	0.00	0.73	–0.02	0.91	–0.02
39	4.27	0.45	3.83	0.51	–0.02	0.76	–0.01	0.70	–0.01	0.87	–0.01
40	4.28	0.50	3.85	0.57	0.01	0.79	–0.03	0.72	0.04	0.91	0.04
41	4.11	0.53	3.71	0.55	–0.01	0.80	–0.01	0.71	0.00	0.87	0.00
42	4.25	0.54	3.74	0.61	0.01	0.74	–0.02	0.73	0.03	0.95	0.03
43	4.21	0.48	3.77	0.53	–0.03	0.77	0.01	0.73	–0.03	0.89	–0.04
44	4.15	0.52	3.71	0.57	0.01	0.80	–0.03	0.73	0.03	0.90	0.04
45	4.15	0.50	3.68	0.56	–0.02	0.79	–0.01	0.70	–0.01	0.87	–0.02
46	4.27	0.50	3.80	0.57	–0.03	0.80	0.01	0.73	–0.04	0.92	–0.04
47	4.26	0.52	3.82	0.57	–0.05	0.82	0.02	0.72	–0.07	0.90	–0.07
48	4.20	0.51	3.78	0.56	–0.03	0.80	0.01	0.71	–0.04	0.88	–0.04
49	4.28	0.52	3.82	0.58	0.02	0.78	–0.03	0.71	0.05	0.90	0.05
50	4.24	0.50	3.80	0.54	–0.05	0.78	0.03	0.68	–0.07	0.87	–0.09
51	4.23	0.52	3.73	0.56	0.01	0.81	–0.03	0.71	0.05	0.88	0.05
52	4.26	0.49	3.83	0.54	–0.01	0.76	–0.01	0.70	0.01	0.88	0.01
53	4.24	0.51	3.82	0.56	0.00	0.80	–0.02	0.69	0.02	0.87	0.02
54	4.25	0.52	3.77	0.55	0.01	0.76	–0.03	0.70	0.04	0.86	0.04
55	4.16	0.51	3.70	0.54	0.01	0.77	–0.03	0.72	0.05	0.88	0.05
56	4.20	0.55	3.75	0.59	–0.07	0.78	0.05	0.70	–0.13	0.91	–0.14
57	4.24	0.50	3.79	0.55	–0.05	0.76	0.02	0.68	–0.07	0.87	–0.08
58	4.17	0.52	3.71	0.58	0.00	0.79	–0.02	0.73	0.03	0.91	0.03
59	4.27	0.50	3.82	0.56	–0.04	0.81	0.02	0.72	–0.07	0.91	–0.07
60	4.24	0.48	3.81	0.54	0.00	0.77	–0.01	0.73	0.01	0.90	0.01
61	4.25	0.46	3.85	0.51	.00	0.74	–0.01	0.69	0.02	0.85	0.02
62	4.28	0.53	3.80	0.59	–0.04	0.76	0.02	0.72	–0.06	0.93	–0.06
63	4.22	0.50	3.78	0.54	–0.03	0.77	0.01	0.69	–0.04	0.85	–0.04
64	4.23	0.54	3.73	0.58	0.06	0.82	–0.09	0.76	0.15	0.94	0.16

(Table continues)

Table D1 (continued)

Prompt	Expected essay scores				Residual (observed-expected)				Mean	Pooled	Residual
	European		East Asian		European		East Asian		resid.	SD	effect
no.	M	SD	M	SD	M	SD	M	SD	diff.	(obs)	size
65	4.25	0.49	3.83	0.53	–0.05	0.78	0.04	0.70	–0.10	0.88	–0.11
66	4.31	0.45	3.92	0.51	0.00	0.75	–0.02	0.70	0.03	0.85	0.03
67	4.21	0.56	3.72	0.59	–0.03	0.80	0.01	0.72	–0.04	0.91	–0.05
68	4.22	0.46	3.84	0.50	0.00	0.75	–0.02	0.70	0.02	0.85	0.03
69	4.26	0.49	3.78	0.55	–0.03	0.76	0.01	0.73	–0.04	0.91	–0.04
70	4.25	0.49	3.77	0.55	–0.04	0.77	0.02	0.71	–0.06	0.89	–0.07
71	4.15	0.45	3.75	0.48	–0.03	0.71	0.01	0.68	–0.03	0.82	–0.04
72	4.17	0.49	3.74	0.54	0.04	0.77	–0.05	0.69	0.09	0.85	0.10
73	4.25	0.53	3.74	0.58	0.02	0.82	–0.05	0.74	0.07	0.92	0.07
74	4.22	0.52	3.79	0.57	0.00	0.78	–0.02	0.69	0.02	0.88	0.03
75	4.14	0.53	3.69	0.57	–0.02	0.79	0.00	0.71	–0.02	0.90	–0.03
76	4.30	0.47	3.91	0.51	0.02	0.73	–0.03	0.66	0.05	0.82	0.06
77	4.35	0.52	3.90	0.55	–0.04	0.77	0.02	0.71	–0.06	0.88	–0.07
78	4.11	0.55	3.64	0.59	0.02	0.76	–0.03	0.71	0.05	0.90	0.06
79	4.19	0.53	3.67	0.61	–0.01	0.77	0.00	0.72	–0.01	0.93	–0.01
80	4.16	0.46	3.74	0.50	–0.03	0.74	0.01	0.70	–0.04	0.85	–0.05
81	4.17	0.54	3.69	0.61	–0.01	0.80	–0.01	0.72	0.00	0.94	0.00
82	4.33	0.48	3.87	0.54	–0.02	0.76	0.00	0.70	–0.02	0.88	–0.03
83	4.14	0.54	3.73	0.56	–0.03	0.75	0.01	0.67	–0.04	0.84	–0.05
84	4.22	0.51	3.83	0.53	0.00	0.76	–0.02	0.66	0.02	0.81	0.03
85	4.20	0.48	3.78	0.51	–0.01	0.76	–0.01	0.69	0.01	0.85	0.01
86	4.23	0.50	3.78	0.56	0.04	0.73	–0.05	0.71	0.09	0.90	0.10
87	4.19	0.48	3.74	0.53	–0.05	0.79	0.02	0.70	–0.07	0.87	–0.08
Total	4.23	0.51	3.79	0.55	–0.01	0.77	–0.01	0.71	0.00	0.88	0.00

Appendix E

Uniform and Nonuniform Effect Sizes

Tables E1 and E2 show uniform and nonuniform effect sizes based on R^2 values for English language ability, native language group, and English language ability by native language group interaction terms from the full (Step 3) logistic regression model:

$$\left(P(U_j \mid x, D) = \frac{1}{1 + \exp[-(\beta_{0i} + \beta_1 x + \beta_2 D_m + \beta_3 x D_m)]} \right)$$

Table E1

Uniform, Nonuniform, and Total R^2 Effect Sizes for 81 Prompts

	R^2 changes						χ^2 test for added terms					
Prompt	R^2 values			R^2 effect size			Ability (A)		Group (G)		A*G	
no.	A	G	A*G	Uni	Non	Total	χ^2	*p*	χ^2	*p*	χ^2	*p*
1	0.3679	0.3693	0.3712	0.0014	0.0019	0.0033	833.92	<.0001	5.33	0.0209	6.44	0.0112
3	0.3930						1,353.62	<.0001				
4	0.4427						1,428.16	<.0001				
5	0.3588						1,412.54	<.0001				
6	0.3932	0.3966	0.3981	0.0034	0.0015	0.0049	1,286.42	<.0001	18.73	<.0001	7.63	0.0057
7	0.3848	0.3858	0.3883	0.0010	0.0025	0.0035	1,224.09	<.0001	5.27	0.0217	12.44	0.0004
8	0.3310						640.30	<.0001				
10	0.3739						1,070.71	<.0001				
11	0.4008	0.4016	0.4026	0.0008	0.0010	0.0018	1,320.38	<.0001	4.13	0.0421	5.52	0.0188
12	0.3464	0.3515		0.0051		0.0051	788.22	<.0001	17.60	<.0001		
13	0.3738						902.09	<.0001				
14	0.4115						999.25	<.0001				
15	0.3587	0.3596	0.361	0.0009	0.0014	0.0023	1,509.25	<.0001	5.67	0.0173	8.90	0.0029
19	0.3487						868.14	<.0001				
21	0.3938						1,209.79	<.0001				
22	0.4579						1,136.08	<.0001				
23	0.3566						1,210.20	<.0001				
24	0.3885						1,211.72	<.0001				
25	0.3758	0.3769		0.0011		0.0011	1,140.01	<.0001	5.23	0.0222		
26	0.4499						1,024.06	<.0001				
27	0.3402	0.3412	0.3425	0.001	0.0013	0.0023	1,133.01	<.0001	5.41	0.0200	6.38	0.0115

(Table continues)

Table E1 (continued)

Prompt	R^2 changes						χ^2test for added terms					
	R^2 values			R^2 effect size			Ability (A)		Group (G)		A*G	
no.	A	G	A*G	Uni	Non	Total	χ^2	*p*	χ^2	*p*	χ^2	*p*
28	0.4142						922.25	<.0001				
29	0.4207						830.38	<.0001				
30	0.3745	0.3769	0.3786	0.0024	0.0017	0.0041	1,127.71	<.0001	11.54	0.0007	7.58	0.0059
31	0.3541						1,022.00	<.0001				
32	0.3323						842.15	<.0001				
33	0.3489						1,420.43	<.0001				
34	0.3612						1,305.63	<.0001				
35	0.4020						1,263.11	<.0001				
36	0.3606	0.3622	0.3636	0.0016	0.0014	0.0030	842.67	<.0001	6.07	0.0137	5.13	0.0236
37	0.3727						1,018.43	<.0001				
38	0.3722						1,472.36	<.0001				
39	0.3425						911.48	<.0001				
40	0.3648						1,068.91	<.0001				
41	0.3619						1,278.17	<.0001				
42	0.4189						1,406.76	<.0001				
43	0.3497						1,083.47	<.0001				
44	0.3690						1,410.95	<.0001				
45	0.3765						1,137.47	<.0001				
46	0.3681						1,383.28	<.0001				
47	0.3745	0.3757	0.3785	0.0012	0.0028	0.0040	1,800.78	<.0001	8.99	0.0027	20.64	<.0001
48	0.3648						832.02	<.0001				
49	0.3901	0.3909	0.3919	0.0008	0.0010	0.0018	1,780.66	<.0001	5.28	0.0216	7.68	0.0056
50	0.3733	0.3752	0.3759	0.0019	0.0007	0.0026	1,560.84	<.0001	12.74	0.0004	4.85	0.0276
51	0.3807						1,223.25	<.0001				
52	0.3652						1,217.74	<.0001				
53	0.3663						959.46	<.0001				
54	0.3887						1,043.39	<.0001				
55	0.3679						1,112.90	<.0001				
56	0.3975	0.4018	0.4036	0.0043	0.0018	0.0061	1,112.27	<.0001	20.64	<.0001	8.44	0.0037
57	0.3868	0.3886	0.3908	0.0018	0.0022	0.0040	1,445.98	<.0001	11.08	0.0009	13.41	0.0003
58	0.3782						1,322.37	<.0001				
59	0.3625	0.3634	0.3641	0.0009	0.0007	0.0016	1,543.11	<.0001	6.40	0.0114	4.24	0.0396
60	0.3511						910.12	<.0001				
61	0.3424						1,237.05	<.0001				
62	0.3949						882.01	<.0001				
63	0.3773						1,226.39	<.0001				
64	0.3754	0.3827		0.0073		0.0073	1,045.92	<.0001	32.22	<.0001		

(Table continues)

Table E1 (continued)

Prompt no.	R^2 changes: R^2 values A	G	A*G	R^2 effect size Uni	Non	Total	χ^2 test for added terms: Ability (A) χ^2	p	Group (G) χ^2	p	A*G χ^2	p
65	0.3552	0.3582	0.3593	0.0030	0.0011	0.0041	1,232.78	<.0001	16.15	<.0001	5.77	0.0163
66	0.3446						1,228.68	<.0001				
67	0.3985	0.3991	0.4023	0.0006	0.0032	0.0038	1,372.53	<.0001	3.87	0.0491	17.61	<.0001
68	0.3382						991.90	<.0001				
69	0.3721						1,218.74	<.0001				
70	0.3731	0.3744	0.3764	0.0013	0.0020	0.0033	1,231.70	<.0001	6.77	0.0093	10.49	0.0012
71	0.3467						1,119.80	<.0001				
72	0.3694	0.3721	0.3731	0.0027	0.0010	0.0037	1,641.95	<.0001	18.97	<.0001	7.32	0.0068
73	0.385	0.3861		0.0011		0.0011	1,215.49	<.0001	5.46	0.0195		
74	0.3924						1,190.45	<.0001				
75	0.3898						1,335.88	<.0001				
76	0.3657	0.3668		0.0011		0.0011	949.63	<.0001	4.30	0.0382		
77	0.3829	0.3841	0.3858	0.0012	0.0017	0.0029	1,273.37	<.0001	6.64	0.0100	8.84	0.0030
78	0.4124	0.4137		0.0013		0.0013	1,082.46	<.0001	5.93	0.0149		
79	0.4137						1,081.43	<.0001				
80	0.3453						742.05	<.0001				
81	0.3945						1,110.51	<.0001				
82	0.3726						1,171.43	<.0001				
83	0.4034						804.10	<.0001				
84	0.377						712.90	<.0001				
85	0.3572						821.33	<.0001				
86	0.3887	0.3914		0.0027		0.0027	801.06	<.0001	8.82	0.0030		
87	0.3594	0.3610	0.3618	0.0016	0.0008	0.0024	1,498.64	<.0001	10.24	0.0014	5.27	0.0217

Table E2

Intercept and Slope Parameters for Logistic Regression for 81 Prompts

Prompt	Intercepts										Slopes		
no.	β_{01}	β_{02}	β_{03}	β_{04}	β_{05}	β_{06}	β_{07}	β_{08}	β_{09}	β_{10}	β_1 (A)	β_2 (G)	β_3 (A*G)
1	–6.50	–5.70	–4.05	–3.10	–1.65	–0.67	0.76	1.79	2.84	3.91	–0.59	–0.21	0.08
3	–6.15	–5.30	–4.09	–3.09	–1.71	–0.61	0.66	1.65	2.59	3.91	–0.53		
4	–6.81	–5.50	–4.36	–3.24	–1.69	–0.62	0.74	1.87	2.88	4.24	–0.59		
5	–6.99	–5.97	–4.34	–3.22	–1.80	–0.67	0.78	1.86	2.78	4.16	–0.51		
7	–6.54	–5.51	–4.18	–3.12	–1.70	–0.62	0.70	1.78	2.73	3.87	–0.56	0.27	0.07
8	–6.71	–5.78	–4.68	–3.58	–2.08	–0.94	0.55	1.61	2.62	3.90	–0.58	0.16	0.09
9	–7.29	–6.18	–4.54	–3.58	–2.07	–0.99	0.46	1.53	2.65	3.80	–0.48		
10	–5.91	–5.05	–3.87	–2.81	–1.48	–0.31	0.96	1.94	2.91	4.19	–0.53		
11	–6.24	–5.22	–3.95	–2.95	–1.60	–0.50	0.92	2.00	3.05	4.34	–0.59	–0.14	0.06
12	–7.09	–6.07	–4.50	–3.47	–1.84	–0.81	0.72	1.72	2.79	3.93	–0.48	0.35	
13	–6.90	–5.94	–4.53	–3.67	–2.19	–1.05	0.64	1.71	2.92	4.13	–0.52		
14	–6.30	–5.38	–4.26	–3.17	–1.73	–0.62	0.88	1.96	2.96	4.14	–0.56		
15	–6.22	–5.43	–4.01	–3.18	–1.68	–0.60	0.89	1.95	2.95	4.20	–0.53	0.13	0.07
19	–6.32	–5.84	–4.60	–3.47	–1.97	–0.93	0.55	1.53	2.56	3.75	–0.48		
21	–6.92	–5.72	–4.23	–3.18	–1.71	–0.58	0.80	1.79	2.79	4.08	–0.54		
22	–6.30	–5.44	–4.29	–3.21	–1.63	–0.53	0.94	2.00	2.96	4.09	–0.60		
23	–6.73	–5.79	–4.08	–3.10	–1.65	–0.53	0.91	1.88	2.82	4.20	–0.50		
24	–6.94	–5.93	–4.42	–3.36	–1.91	–0.82	0.63	1.65	2.68	3.89	–0.53		
25	–6.81	–5.97	–4.73	–3.64	–2.02	–0.88	0.58	1.69	2.71	3.87	–0.51	0.16	
26	–6.26	–5.46	–4.36	–3.25	–1.74	–0.75	0.71	1.78	2.81	3.97	–0.61		
27	–7.15	–6.00	–4.61	–3.60	–1.95	–0.90	0.59	1.70	2.64	3.65	–0.51	0.14	0.06
28	–6.35	–5.28	–4.21	–3.08	–1.61	–0.43	0.85	1.90	2.85	4.32	–0.57		
29	–5.84	–5.18	–3.83	–2.81	–1.46	–0.30	1.01	1.95	2.85	4.01	–0.56		
30	–6.23	–5.33	–3.98	–2.94	–1.49	–0.36	0.92	1.95	2.94	4.08	–0.58	–0.26	0.07
31	–6.39	–5.66	–4.52	–3.43	–1.89	–0.84	0.57	1.54	2.49	3.65	–0.51		
32	–6.50	–5.58	–4.58	–3.59	–2.12	–0.93	0.69	1.72	2.81	4.12	–0.48		
33	–6.52	–5.74	–4.39	–3.38	–1.91	–0.83	0.61	1.65	2.66	3.91	–0.49		
34	–6.97	–5.81	–4.30	–3.27	–1.87	–0.73	0.60	1.67	2.63	3.75	–0.51		
35	–6.96	–5.69	–4.43	–3.31	–1.85	–0.73	0.61	1.76	2.80	4.09	–0.56		
36	–6.68	–5.82	–4.35	–3.33	–1.86	–0.72	0.65	1.74	2.82	4.12	–0.54	0.18	0.07
37	–7.59	–5.83	–4.14	–3.15	–1.79	–0.66	0.74	1.83	2.82	4.03	–0.51		
38	–6.42	–5.54	–4.32	–3.21	–1.90	–0.87	0.56	1.65	2.62	3.85	–0.51		
39	–7.19	–6.05	–4.33	–3.25	–1.79	–0.65	0.69	1.79	2.82	4.03	–0.48		
40	–6.63	–5.67	–4.26	–3.13	–1.70	–0.61	0.67	1.70	2.61	3.81	–0.51		
41	–6.23	–5.33	–3.84	–2.81	–1.42	–0.33	0.94	2.01	3.04	4.18	–0.52		
42	–6.42	–5.42	–4.09	–3.16	–1.64	–0.52	0.81	1.91	2.85	4.19	–0.57		

(Table continues)

Table E2 (continued)

Prompt	Intercepts										Slopes		
no.	β_{01}	β_{02}	β_{03}	β_{04}	β_{05}	β_{06}	β_{07}	β_{08}	β_{09}	β_{10}	β_1 (A)	β_2 (G)	β_3 (A*G)
43	–5.93	–5.38	–4.12	–3.16	–1.69	–0.59	0.76	1.83	2.84	4.02	–0.50		
44	–6.14	–5.24	–3.96	–2.92	–1.54	–0.50	0.85	1.87	2.84	3.92	–0.51		
45	–5.91	–5.15	–4.07	–3.07	–1.63	–0.42	0.90	1.95	2.92	4.17	–0.52		
46	–6.30	–5.43	–4.17	–3.06	–1.75	–0.63	0.69	1.72	2.62	3.82	–0.51		
47	–6.04	–5.12	–3.93	–3.06	–1.64	–0.53	0.80	1.82	2.78	4.01	–0.58	–0.19	0.10
48	–6.72	–5.36	–4.05	–3.05	–1.74	–0.70	0.71	1.70	2.80	4.06	–0.50		
49	–6.55	–5.72	–4.29	–3.31	–1.81	–0.72	0.61	1.69	2.69	3.79	–0.57	0.11	0.06
50	–6.45	–5.59	–4.20	–3.24	–1.64	–0.55	0.88	1.98	2.92	4.14	–0.55	–0.22	0.05
51	–6.39	–5.41	–4.09	–3.05	–1.61	–0.53	0.86	1.84	2.74	3.93	–0.52		
52	–6.38	–5.60	–4.42	–3.29	–1.77	–0.65	0.70	1.79	2.76	4.07	–0.51		
53	–6.19	–5.37	–4.16	–3.27	–1.76	–0.67	0.66	1.72	2.78	3.89	–0.52		
54	–6.71	–5.71	–4.30	–3.24	–1.70	–0.60	0.78	1.86	2.83	3.95	–0.55		
55	–6.67	–5.52	–3.99	–3.02	–1.51	–0.45	0.89	1.92	2.87	4.17	–0.51		
56	–5.74	–5.15	–3.98	–2.94	–1.49	–0.46	0.98	2.04	3.16	4.44	–0.62	–0.34	0.08
57	–6.37	–5.51	–4.14	–3.16	–1.62	–0.48	0.95	2.06	3.07	4.30	–0.60	–0.24	0.09
58	–5.85	–5.17	–3.99	–2.94	–1.52	–0.50	0.82	1.91	2.87	4.09	–0.52		
59	–6.22	–5.31	–4.12	–3.22	–1.63	–0.61	0.68	1.78	2.71	3.90	–0.54	–0.16	0.05
60	–6.42	–5.66	–4.18	–3.21	–1.81	–0.77	0.59	1.64	2.75	3.91	–0.49		
61	–6.67	–5.82	–4.35	–3.27	–1.90	–0.81	0.66	1.82	2.82	4.09	–0.48		
62	–6.11	–5.40	–4.29	–3.23	–1.86	–0.72	0.60	1.74	2.79	3.92	–0.54		
63	–6.42	–5.59	–4.29	–3.41	–1.85	–0.72	0.76	1.79	2.86	4.04	–0.51		
64	–5.95	–5.33	–4.16	–3.09	–1.75	–0.68	0.49	1.51	2.42	3.69	–0.48	0.42	
65	–6.25	–5.41	–4.12	–3.05	–1.62	–0.49	0.80	1.92	3.00	4.17	–0.55	–0.28	0.06
66	–6.36	–5.86	–4.63	–3.64	–2.09	–0.92	0.56	1.57	2.73	3.92	–0.49		
67	–5.91	–5.04	–3.94	–2.83	–1.47	–0.30	1.00	1.98	2.89	4.07	–0.61	–0.15	0.11
68	–6.84	–6.00	–4.33	–3.27	–1.77	–0.66	0.80	1.80	2.90	4.07	–0.50		
69	–7.09	–5.73	–4.20	–3.15	–1.79	–0.68	0.71	1.76	2.70	3.95	–0.51		
70	–6.20	–5.45	–3.99	–3.04	–1.54	–0.46	0.93	2.00	3.01	4.30	–0.59	–0.21	0.09
71	–6.94	–5.73	–4.42	–3.36	–1.82	–0.60	0.91	1.99	3.10	4.34	–0.49		
72	–6.66	–5.67	–4.27	–3.22	–1.74	–0.62	0.82	1.92	2.90	4.01	–0.53	0.24	0.06
73	–7.12	–5.60	–4.10	–3.04	–1.65	–0.59	0.70	1.69	2.53	3.64	–0.50	0.16	
74	–6.33	–5.84	–4.32	–3.19	–1.71	–0.56	0.89	1.87	2.77	3.90	–0.55		
75	–6.07	–5.24	–4.03	–3.04	–1.54	–0.42	0.87	2.01	2.95	4.06	–0.54		
76	–8.35	–6.94	–4.90	–3.80	–2.15	–1.09	0.55	1.60	2.72	3.84	–0.50	0.16	
77	–6.39	–5.69	–4.29	–3.31	–1.87	–0.75	0.64	1.70	2.77	3.93	–0.58	–0.19	0.08
78	–6.33	–5.48	–3.98	–3.05	–1.69	–0.52	0.94	1.99	2.94	4.35	–0.55	0.18	
79	–6.45	–5.28	–3.96	–2.92	–1.52	–0.39	0.96	2.02	3.01	4.21	–0.56		
80	–6.36	–5.36	–4.17	–3.31	–1.69	–0.54	0.88	1.90	3.01	4.42	–0.49		

(Table continues)

Table E2 (continued)

Prompt	Intercepts										Slopes		
no.	β_{01}	β_{02}	β_{03}	β_{04}	β_{05}	β_{06}	β_{07}	β_{08}	β_{09}	β_{10}	β_1 (A)	β_2 (G)	β_3 (A*G)
81	–6.34	–5.25	–3.76	–2.74	–1.46	–0.35	0.91	1.95	2.96	4.23	–0.55		
82	–6.91	–5.81	–4.54	–3.51	–2.01	–0.84	0.59	1.61	2.65	3.84	–0.51		
83	–6.97	–5.86	–4.28	–3.26	–1.62	–0.52	1.00	2.12	3.09	4.25	–0.56		
84	–6.54	–6.03	–4.42	–3.33	–1.83	–0.80	0.80	1.82	3.00	4.30	–0.53		
85	–6.94	–5.83	–4.26	–3.40	–1.74	–0.75	0.82	1.92	2.92	4.15	–0.50		
86	–7.31	–6.04	–4.51	–3.40	–1.78	–0.80	0.67	1.75	2.80	4.05	–0.52	0.26	
87	–6.18	–5.43	–4.02	–2.90	–1.58	–0.49	0.93	2.04	3.09	4.19	–0.55	–0.20	0.05

Appendix F

Scoring Rubrics for TOEFL CBT Writing Prompts

The content of this appendix is excerpted from the *Computer-based TOEFL Test Score User Guide* (ETS, 1998, 1999).

6 An essay at this level

- effectively addresses the writing task
- is well organized and well developed
- uses clearly appropriate details to support a thesis or illustrate ideas
- displays consistent facility in the use of language
- demonstrates syntactic variety and appropriate word choice, though it may have occasional errors

5 An essay at this level

- may address some parts of the task more effectively than others
- is generally well organized and well developed
- uses details to support a thesis or illustrate an idea
- displays facility in the use of the language
- demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors

4 An essay at this level

- addresses the writing topic, but slight parts of the task
- is adequately organized and developed
- uses some details to support a thesis or illustrate an idea
- displays adequate but possibly inconsistent facility with syntax and use
- may contain some errors that occasionally obscure meaning

3 An essay at this level may reveal one or more of the following weaknesses:

- inadequate organization or development
- inappropriate or insufficient details to support or illustrate generalizations
- a noticeably inappropriate choice of words or word forms
- an accumulation of errors in sentence structure and/or usage

2 An essay at this level is seriously flawed by one or more of the following weaknesses

- serious disorganization or underdevelopment
- little or no detail, or irrelevant specifics
- serious and frequent errors in sentence structure or usage
- serious problems with focus

1 An essay at this level

- may be incoherent
- may be underdeveloped
- may contain severe and persistent writing errors

0 An essay will be rated 0 if it

- contains no response
- merely copies the topic
- is off-topic
- is written in a foreign language
- consists only of keystroke characters

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

Email: toefl@ets.org

Web site: www.ets.org/toefl

* America Samoa, Guam, Puerto Rico, and US Virgin Islands

I.N. 725615