TOEFL

Research Reports

Report 76
February 2004

An Analysis of TOEFL CBT Writing Prompt Difficulty and Comparability for Different Gender Groups

Hunter Breland

Yong-Won Lee

Michelle Najarian

Eiji Muraki

An Analysis of TOEFL CBT Writing Prompt Difficulty and Comparability for Different Gender Groups

Hunter Breland, Yong-Won Lee, and Michelle Najarian
ETS, Princeton, NJ

and

Eiji Muraki
Tohoku University, Sendai, Japan

RR-04-05

ETS is an Equal Opportunity/Affirmative Action Employer.

Abstract

This investigation of the comparability of writing assessment prompts was conducted in two phases. In an exploratory Phase I, 47 writing prompts administered in the computer-based Test of English as a Foreign Language™ (TOEFL® CBT) from July through December 1998 were examined. Logistic regression procedures were used to estimate prompt difficulty and gender effects. A panel of experts reviewed selected prompts, and a taxonomy of prompt characteristics was developed and related to prompt difficulty and gender differences. In Phase II, 87 prompts administered from July 1998 through March 2000 were analyzed. All of the prompts used in Phase I, together with 40 new prompts, were analyzed using the larger Phase II database. Recommendations are made for statistical quality control procedures to identify less comparable prompts.

Key words: computer-based writing assessment, essay prompts, comparability, fairness, polytomous DIF (differential item functioning), gender, 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 under the direction of 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 meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. 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	University of Auckland
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.toefl.org

Acknowledgements

Several members of the ETS staff in addition to the authors contributed to this project. Robert Kantor served as a general advisor on a number of matters, convened panels of experts to review writing prompts identified in the analyses, served as a reviewer himself, and provided the TOEFL prompt classifications used in the taxonomy of prompt characteristics. Guinn Roberts also served as a member of the review panels. Dan Tumposky assisted with the prompt taxonomy development, as did Diane Gruenberg (of The College of New Jersey). Corinne Reslier provided the TOEFL data tape used for the analyses and explained numerous details of the data, and Tom Jirele advised on some of the data analyses for the Phase I Study. Youn-Hee Lim, Ting Lu, and Cindy Nguyen helped us with data preparation and analyses for the Phase II study. We appreciate also the assistance provided by the external members of the review panels: Lynn Goldstein of the Monterey Institute of International Studies, Joan Carson of Georgia State University, Terry Santos of Humboldt State University, and Alistair Cumming of the Ontario Institute for Studies in Education, University of Toronto. Thanks also are due to external reviewers Lyle Bachman and Brenda Lim, who provided useful comments and suggestions.

Table of Contents

List of Tables

List of Figures

Introduction

The focus of the present investigation was on the comparability of the prompts used in writing skill assessment. It is important to examine the comparability of prompts in the computer-based Test of English as a Foreign Language™ (TOEFL® CBT) because each examinee receives only a single prompt, and this prompt is generally not the same for all examinees. If the prompts are not comparable in difficulty, those examinees receiving the most difficult prompts would be disadvantaged and those receiving the least difficult prompts would be advantaged. To our knowledge, no statistical method exists to control for difficulty when only one item or prompt is administered to each examinee, and yet it is important in testing programs to ensure that all examinees are administered tests of equivalent difficulty. For these reasons, Stansfield and Ross (1988), in their long-term research agenda for TOEFL writing assessment, stated that the highest priority should be given to the issue of comparability of scores obtained for different writing prompts. The comparability of prompts for different groups of examinees, such as gender groups, is also of importance.

Gender differences on free-response writing examinations have tended to favor females, but the magnitude of gender differences varies across populations of examinees. For example, the National Assessment of Educational Progress (1994) has reported gender differences in performance on essay tests for national random samples exceeding one-half of a standard deviation in grades 8 and 12. Similar results have been reported for some statewide examinations at grade 8 (Englehard, Gordon, & Gabrielson, 1991). In college-bound populations, gender differences in performance on essay tests of writing skill have been much smaller, ranging from a little over one-tenth to about one-third of a standard deviation, but still favoring females (Breland & Griswold, 1982; Breland & Jones, 1982; Bridgeman & Bonner, 1994). In graduate school applicant populations, females have averaged about one-tenth of a standard deviation higher on essay tests than males (Bridgeman & McHale, 1996: Schaeffer, Briel, & Fowles, 2001).

Females also tend to score slightly higher than males on writing tests in populations for whom English is a second language. In random samples of Test of English as a Foreign Language (TOEFL) examinees, female scores on an essay test averaged about one-tenth of a standard deviation higher than those for males (Golub-Smith, Reese, & Steinhaus, 1993). Two additional studies of ESL students have yielded higher scores for ESL females at the elementary school level. Bermudez and Prater (1994) found that essays written by Hispanic females showed a greater degree of elaboration and received higher holistic scores. Heck and Crislip (2001) studied third grade students in Hawaii and found that females scored higher on both direct and indirect measures of writing skill than did males.

It is not uncommon in psychology, and education more generally, to observe that females perform better on free-response tasks. In a paper on verbal fluency differences in school-age children, Sincoff and Sternberg (1988) reported that girls, especially those above age 11, scored higher than boys on verbal fluency tasks. In a summary paper, Sincoff and Sternberg (1987) discuss two types of verbal ability: verbal fluency and verbal comprehension. Verbal fluency is needed primarily for writing and speaking tasks, while verbal comprehension is needed primarily for reading and listening tasks. Berninger and Fuller (1992) studied written compositions of first, second, and third grade students and found that boys were at greater risk for writing disabilities.

In the Golub-Smith et al. study, the comparability of prompts used for the Test of Written English™ (TWE®) was examined. Eight different prompts were spiraled (that is, administered in a random or near-random manner worldwide at the October 1989 TOEFL administration, with each prompt eliciting approximately 10,000 essays. The results of the analyses conducted indicated small differences in mean scores obtained from some of the prompts, but the investigators had difficulty making definitive statements regarding the meaningfulness of the observed differences. While many of the observed differences in means were so small as to be of no practical significance, differences observed across prompts in the number of examinees at each score level were not. The study suggested that these score distribution differences may warrant further investigation.

Other testing programs have also conducted studies of prompt differences. Pomplun, Wright, Oleka, and Sudlow (1992) studied prompts used for the College Board’s English Composition Test (ECT) with Essay. This study used ECT data for seven prompts administered during the years 1983 to 1990. Differential difficulty was explored through linear regressions of essay scores on objective scores for different sex, language, and ethnic groups. The results of these analyses indicated that, generally, the regressions were consistent across years but that two of the seven prompts studied contained characteristics that may have been related to differential performance. In one of the two identified prompts, the topic of heroes and values may have favored groups more familiar with cultural values. In the other prompt identified, the combination of an abstract topic with an ironic tone may have caused differential performance for those with lower language skills. Further study was recommended of the nature of essay performance of minority and ESL groups.

Although their primary objective was not the study of prompt comparability, other studies have yielded results that are informative concerning the differential difficulty of prompts or the testing of foreign-language populations more generally. Mazzeo, Schmitt, and Bleistein (1993) compared the performance of gender and ethnic groups on the essay and multiple-choice components of Advanced Placement examinations. The results suggested that topic variability may have a greater effect than the variability associated with particular question types or broadly defined content areas. Questions based on passages related to topics such as patriotism, space satellites, and the ruggedness of the American prairie produced the largest group differences, which favored males.

In a comprehensive review of measurement issues related to gender, Willingham and Cole (1997) noted that the specific topic of an essay assessment may affect the performance of different genders. For example, on the Advanced Placement English Language and Composition examination, some topics seemed to favor males, while others favored females. White women performed better than white men on a question that required an evaluation of an assertion about human nature. White men performed better than white women on a topic that asked them to compare the styles of passages written by Native Americans about the harshness of the American prairie (p. 191).

While these previous studies have contributed to an understanding of problems in the assessment of English language writing, they have all been limited by the availability of prompts and by sampling restrictions. Prior to the introduction of computer-based testing, a single prompt was often associated with a single test administration date for TOEFL. There were thus unavoidable confoundings of prompts and samples, which made the comparison of prompts difficult at best. With the new TOEFL CBT administrations, numerous prompts are administered in a random (or near random) fashion to widely varying populations. These new CBT administrations thus offer an opportunity to examine prompt comparability with a rigor that has heretofore been impossible.

Analyses of this type could have important implications for test development. If distinctive patterns are observed for different prompts, these patterns could guide prompt development and selection. An ultimate outcome could be greater efficiency in the development of prompts, and this could contribute to a higher success rate for prompts during pretesting. Or, the ultimate outcome could be a restriction on the number of prompts generated. The analyses would also indicate what criteria might be used to determine whether prompts should be pretested or whether they should be removed from the active pool of prompts. Finally, if different patterns are observed at various score levels across prompts, these patterns can be used to suggest further evaluation and possible improvement of sample papers, annotations, and other topic support material.

The present investigation focused on TOEFL CBT administrations that began in the summer of 1998. These administrations included multiple-choice tests of reading, listening, and structure plus a free-response test of writing, which consists of a brief essay assessment with either word-processed or handwritten response modes. The prompts for the CBT essay are selected for each examinee from a pool of prompts using a complex computer algorithm that begins with a random starting position in a list of essay prompts, giving each prompt in the pool of prompts an even exposure. Because each examinee receives a different question, it is important that the prompts be of reasonably equivalent difficulty. Moreover, questions arise as to whether the prompts are of equivalent difficulty for subgroups of examinees, such as those identified by gender or other categories. The objective of the present investigation was to compare the difficulty of CBT essay prompts for groups of examinees receiving different prompts and for different gender groups receiving the same prompt.

Methods

Instruments

The data available for analysis were from TOEFL administrations conducted between July 1998 and August 2000. The data included scores for TOEFL Reading (linear), Listening (adaptive), Structure (adaptive), and Writing (essay).[1]

The following records were excluded from the database:

1. Records with administrative dates preceding July 24, 1998.
2. Records with a “nonstandard” indicator (*e.g.*, examinees receiving extra time for testing or not taking the listening section).

3. Records with a "special conditions" indicator (*e.g.*, examinees who tested under any conditions that differ from the general test taker population, such as using the Braille test, a large print test, a sign language interpreter).
4. Records where the essay termination was other than normal (*e.g.,* when an examinee terminated without responding to the essay prompt).
5. Records that did not contain two rater essay scores.

Variables

The following variables were selected from the database:

1. *TOEFL Reading score.* This score is based on a linear multiple-choice test of reading and has a score range from 0 to 30.
2. *TOEFL Listening score.* This score is based on an adaptive multiple-choice test of listening comprehension and has a score range of 0 to 30.
3. *TOEFL Structure score.* This score is based on an adaptive multiple-choice test of English grammar and sentence structure and has a range of 0 to 13.
4. *TOEFL Essay score.* This score ranges from 1 to 6 with possibilities of .5 intervals and is based on two independent readings and holistic ratings of the essay response on a 1 to 6 scale (See Appendix G for scoring rubrics). 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.
5. *Gender.*
6. *Prompt identification code.*

In addition to the above variables available from the TOEFL database, the following variables were developed:

7. *Standardized ability, reading.* This is a standardization of Variable 1, with a mean of zero and a standard deviation of 1.0.
8. *Standardized ability, listening.* This is a standardization of Variable 2, with a mean of zero and a standard deviation of 1.0.

9. *Standardized ability, structure.* This is a standardization of Variable 3, with a mean of zero and a standard deviation of 1.0.
10. *English language ability (ELA).* This is the simple sum of Variables 7, 8, and 9.

Samples

The Phase I sample of data analyzed consisted of TOEFL CBT essay data collected from the July 1998 through March 1999 administrations. The total sample included 69,201 females and 79,963 males representing 221 different native countries and 145 different native languages. A total of 1,201 essays without gender identification were excluded from the analysis. Of the 150,364 essays written, 77,390 were word-processed and 72,974 were handwritten. Of 47 essay prompts used for the logistic regression analyses, 35 were introduced in July 1998 and 12 were introduced into the essay pool in October 1998. The total number of examinees for each essay prompt in the logistic regression analyses ranged from 2,053 to 4,314.

In Phase II, the data analyzed were based on all test administrations conducted between July 1998 and August 2000. A total of 5,660 cases that did not provide gender were excluded from the analysis. In total, 632,246 essays written in response to 87 different prompts were included in the analysis. Of 632,246 essays, a total of 336,153 of these examinees were male and 296,093 were female. Sample sizes for both male and female groups were higher than 1,000 for each of the 87 prompts. The total number of examinees for each essay prompt ranged from 2,066 to 11,760. (See Table B1 for more detailed information.)

Logistic Regression Analyses

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 (female and male) groups in the context of this TOEFL CBT writing prompt investigation.

Logistic regression has 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 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 gender dummy group variable (male=0; female=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., male vs. female) variable was entered ($g_i(x, D) = \beta_{0i} + \beta_1 x + \beta_2 D_m$);
- In Step 3, the interaction term (i.e., ELA-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 of the size of R^2 coefficients.

To gauge the amount of the group differences (if any), three different kinds of the effects sizes from the logistic regression were used in this study: (1) the residual-based effect size; (2) R^2 combined with p-values for the χ^2 test and slope parameters; and (3) 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) in the regression model. Expected essay scores for individual examinees' ELA scores were computed from the Step 1 model ($g_i(x, D) = \beta_{0i} + \beta_1 x$). Residual scores were obtained for each individual examinee by subtracting their ELA-predicted essay scores from their raw essay scores, and these residual scores were averaged separately for each gender 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 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 gender group variable into the ability-only regression model (Step 1), and the nonuniform effect size is 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. To gauge the magnitude of effect sizes, we have used suggestions and recommendations from the DIF literature, although the logistic regression procedures used here are not traditional DIF procedures. For DIF analyses, Zumbo (1999) has suggested that, for an item to be classified as displaying DIF (i.e., an aggregate of uniform and nonuniform DIF), the 2-degree of freedom χ^2 test between Step 1 and 3 needs to have a p-value less than or equal to 0.01 and the R^2 difference between them should be at least 0.13 for the essay prompt. Zumbo's DIF classification scheme has been questioned by Jodoin and Gierl (2001), however, who recommended R^2 values of 0.035 (for negligible DIF), 0.035 to 0.070 (for moderate DIF), and greater than 0.070 (for large DIF). Note that these recent DIF thresholds are different from the more established thresholds suggested by Cohen (1988, 1992) of R^2 values of 0.02, 0.13, and 0.26 for "small," "medium," and "large" effect sizes, respectively. The Cohen thresholds for R^2 effect sizes have also been linked to group mean score differences of 0.20, 0.50, and 0.80 in standard deviation units, which have been used when working with differences measured in standard deviation units. Given the variety of classification schemes recommended, some judgment is required in interpreting results.

The group-specific expected score curves were next obtained for those prompts that were flagged because of significant group effects, as explained in Appendix A. 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 two 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 gender 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 examinees of the same English language ability but from different gender groups.

Expert Review of Selected Prompts (Phase I)

A number of essay prompts at extremes of difficulty, gender differences, and distributional differences were sampled for review by an expert panel. In recognition of the special characteristics of ESL writing and its evaluation (see Cumming, 1990a, 1990b; and Cumming & Mellow, 1996), the expert panel included four applied linguists who were ESL experts and two members of the ETS test development staff who had worked on both the TWE and the TOEFL essay. Three of the applied linguists had served as TOEFL readers, and one had been a chief reader. For their review, the panel was provided with two essay responses at each score level for each of the prompts reviewed.

Taxonomy Development (Phase I)

A taxonomy of TOEFL essay prompt characteristics was developed from the expert review results, the logistic regression analyses, the analyses of gender differences, prompt word counts, low frequency word counts, and the TOEFL program prompt classification system. Word frequency, which is related to word difficulty, was analyzed using an electronic word frequency list developed by Breland and Jenkins (1997).

Results

Phase I Results

Since the Phase II sample included all of the data used in Phase I, only those parts of Phase I that were not repeated in Phase II are reported. The two parts of Phase I that were not repeated in Phase II were the expert reviews of selected prompts and the taxonomy of writing prompts developed in Phase I.

Expert review of selected prompts. The panel of six experts reviewed prompts selected on the basis of their difficulty, gender differences, and unusual score distributions. For each of the prompts selected for review, the experts were provided with two examinee responses at each score level and asked to speculate on reasons why prompts may have been more or less difficult.

For the most difficult prompts, the panel speculated that the primary reason for the lower scores may have been that the examinees had little personal experience with the specific topics of the prompts. Special knowledge, some historical, was at times helpful, but examinees tended to use hypothetical examples based on conjecture. One prompt was hypothesized to be outside of the cultural norm for some countries and, as a result, some examinees may not have had any personal experience with the topic. It should be noted, however, that all of the prompts studied had been pretested and had been administered operationally in paper and pencil form. Another speculation was that some examinees might feel uncomfortable responding to certain topics.

The review panel viewed the easy prompts as being opposite to the difficult items: examinees used personal experience, showed that the content was easily accessible to them, demonstrated rich thoughts and remembrances, and were able to describe and analyze these personal experiences.

For the prompts with large gender differences, the panel hypothesized that, on average, women: (1) are more interested in arts and music, (2) care more about housing and living conditions, and (3) are more aware of and interested in the intricacies of human relationships.

The review panel had fewer comments on the prompts with distributional differences, but they tended to offer some of the same arguments concerning topic accessibility as were used for the difficult and easy prompts. There was the feeling that examinees either could or could not get easily into the topic using their personal experience, which would tend to explain fewer scores in the middle than for other prompts. No explanation was offered for why two of the selected prompts had more scores at the mode than other prompts, but this may have resulted from a central tendency for these prompts on the part of readers (perhaps because of a reluctance to assign low or high scores to responses to this prompt).

In a second meeting of experts, this time those with extensive TWE reading experience, three sets of three prompts were reviewed. Three easy, three difficult, and three average prompts were identified for this second reading, and these prompts were mutually exclusive of those used for the first expert review. For this review, the experts were not provided with examinee responses but reviewed the prompts only. These experts found that the easy prompts had in common immediate "hooks" that could be tied to personal experience, as well as an implied organization that writers could take advantage of. The prompts identified as difficult uniformly required more discussion of and construction of a somewhat more abstract paradigm about how the world works. Writers would be less likely to hook into any personal experience or thoughts they may have had prior to seeing these topics. The experts felt that the topics judged as medium in difficulty were ones that writers, while not having necessarily had full personal experience, could easily bring to bear details and examples they had seen in the news, or perhaps even experienced or knew others who had had experience.

These observations of prompt differences are similar to those made by Powers and Fowles (1999) in a study of test takers' judgments of essay prompts. In this study, essay prompts being considered for possible use in a graduate admissions writing test were evaluated by Graduate Record Examinations® test takers. The study identified several features of essay prompts that examinees considered to be important. The "best" essay prompts were judged to be those that they were familiar with and drew on their personal experiences, knowledge, or observations.

Taxonomy of writing prompts. Based on the expert panel review and other information collected in the project, a taxonomy of TOEFL writing prompt characteristics was developed. The 47 prompts analyzed were first ranked in order of difficulty from least difficult to most difficult. Residual-based effect sizes (RBES) from the Step 1 logistic regression analysis were used to indicate gender differences. The general topic of the prompt was indicated, the number of words in the prompt, and the number of low frequency (perhaps more difficult) words in the prompt. A judgment was made of the probability that most examinees would have had personal experiences such that responding to the prompt could be less difficult. Finally, the TOEFL program classification of prompts was included in the taxonomy.

The less difficult prompts were about topics that most examinees should have been familiar with and thus for which the probability of having had a personal experience is generally high. The opposite tended to be true for the most difficult prompts. The prompts having the largest gender differences tended to be about topics such as art and music, roommates, housing, friends, and children. The smallest gender differences tended to be associated with topics such as research, space travel, factories, and advertising.

There appeared to be no relationship between the number of words in a prompt and its difficulty or gender difference, and there was little relationship between the occurrence of low frequency words and prompt difficulty or gender differences. Similarly there appeared to be little if any relationship between the TOEFL classification of prompt types and difficulty or gender. The taxonomy developed is included in this report as Appendix F.

Phase II Results

Analyses of data aggregated across prompts. Table 1 gives the overall means, standard deviations, and standardized mean differences observed for male and female examinees when all 87 prompts were aggregated. The standardized mean ELA difference between the gender groups (–0.01) is not statistically significant even with the large number of cases used. The standardized mean difference in essay scores observed (–0.13) is statistically significant, but it would be viewed as a very small effect size using Cohen's (1988) standard of .20 as "small." These results are similar to those observed in the previous study of TOEFL writing by Golub-Smith et al. (1993). That study analyzed essay score gender differences in eight random samples of TOEFL examinees and obtained an average standardized mean difference of –.08, favoring females.

When the regression of essay scores on ELA was conducted using the logistic regression procedure, the adjusted gender difference remained the same as the observed gender difference (–0.13). There was no change because the random assignment of prompts to examinees had apparently already controlled for ELA differences.

Table 1

Means and Standard Deviations of English Language Ability and Observed Essay Scores for Male and Female Examinee Groups and Standardized Mean Differences

Variable/Gender	*N*	Mean	SD	*d*
TOEFL essay score:				
Male examinee group	336,153	3.99	0.99	–0.13*
Female examinee group	296,093	4.12	0.94	
English Language ability:				
Male examinee group	336,153	0.01	2.79	–0.01
Female examinee group	296,093	0.03	2.61	

Note. Based on Phase II analyses of 87 prompts. The standardized mean difference, *d*, was computed by subtracting the female mean from the male mean and then dividing by the average standard deviation.
* p < 0.01 two-tailed.

Analyses of mean gender differences for prompts. Figures 1 and 2 show plots of TOEFL English language ability and individual prompt score means. The most obvious observation that can be made from these plots is that the mean English language ability scores (Figure 1) of male and female examinees are almost the same for most of the 87 prompts analyzed. Despite the lack of difference in English language ability between the two groups, however, the mean essay scores for female examinees are slightly higher than those for male examinees (Figure 2).

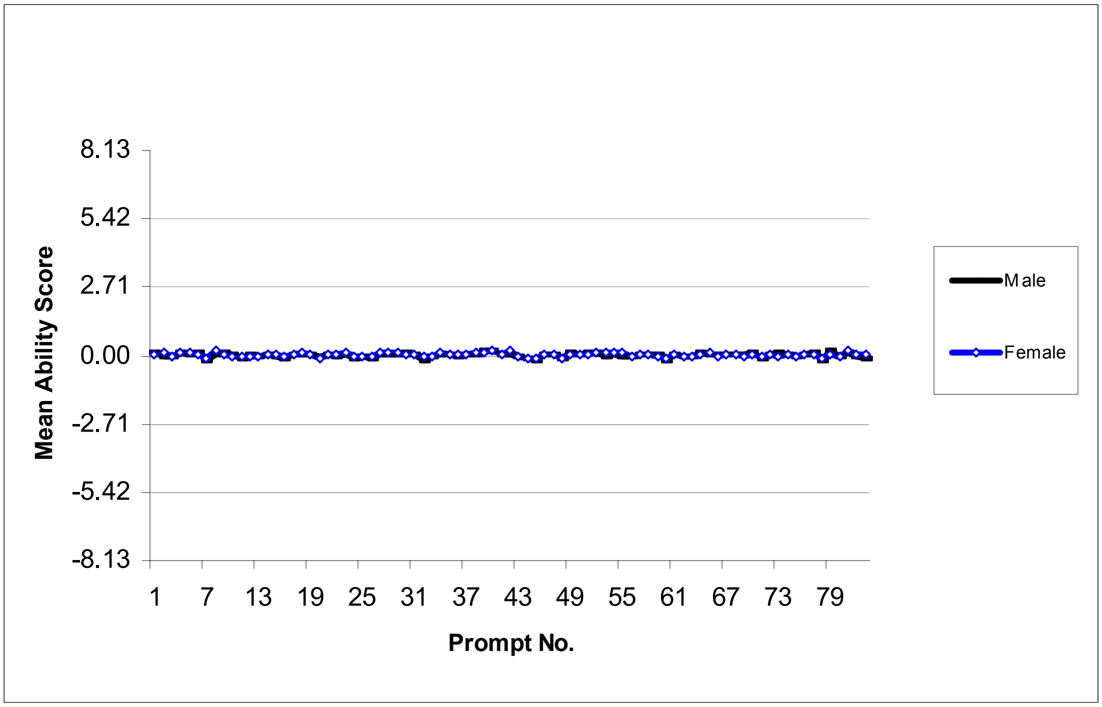

Figure 1. **Mean English language ability (ELA) for male and female examinee groups for 87 Phase II essay prompts.**

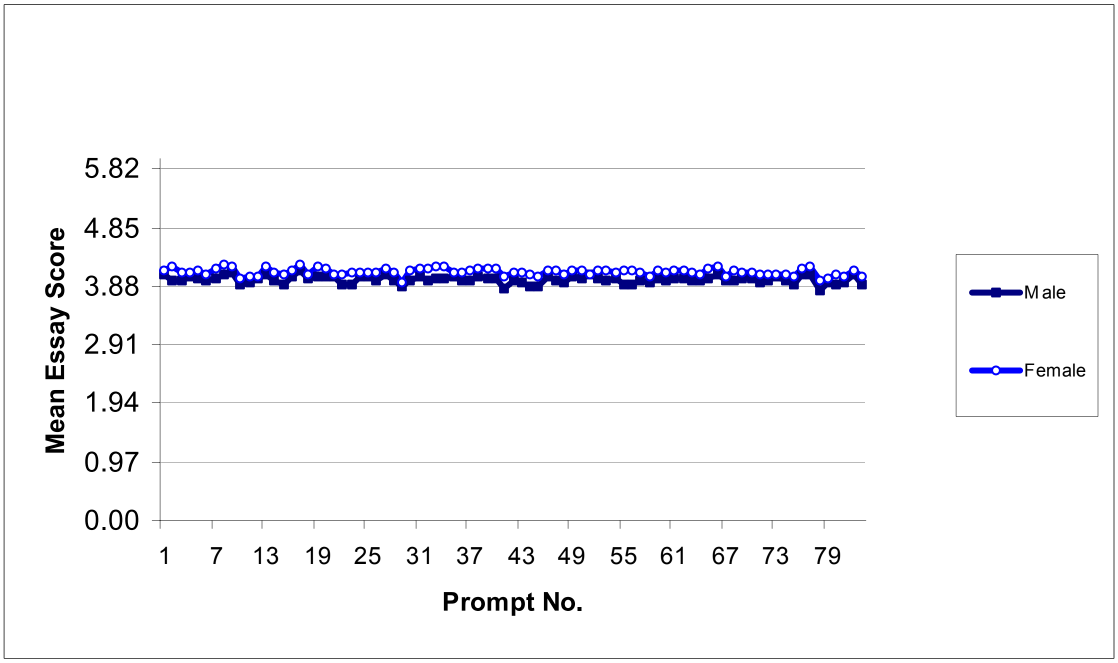

Figure 2. **Mean essay scores for male and female examinee groups for 87 Phase II essay prompts.**

Figures 3 and 4 present the results obtained when the logistic regression procedure was applied at the individual prompt level. Figure 3 shows that female examinees scored slightly higher than expected on most of the 87 prompts, while male examinees scored slightly lower than expected. Figure 4 shows the gender effect sizes for individual prompts after controlling for ELA differences. The negative residual scores for male examinees suggest that female examinees tend to score slightly higher on all the prompts. The residual-based effect sizes ranged from –0.24 to 0.00, with a mean of –0.13. Among them, prompts 2, 33, and 56 had the largest negative effect sizes, which ranged from

–0.21 to –0.24 (favoring female examinees). Although these were the largest gender effect sizes, they would be considered “small” effect sizes by Cohen’s (1988) rule.

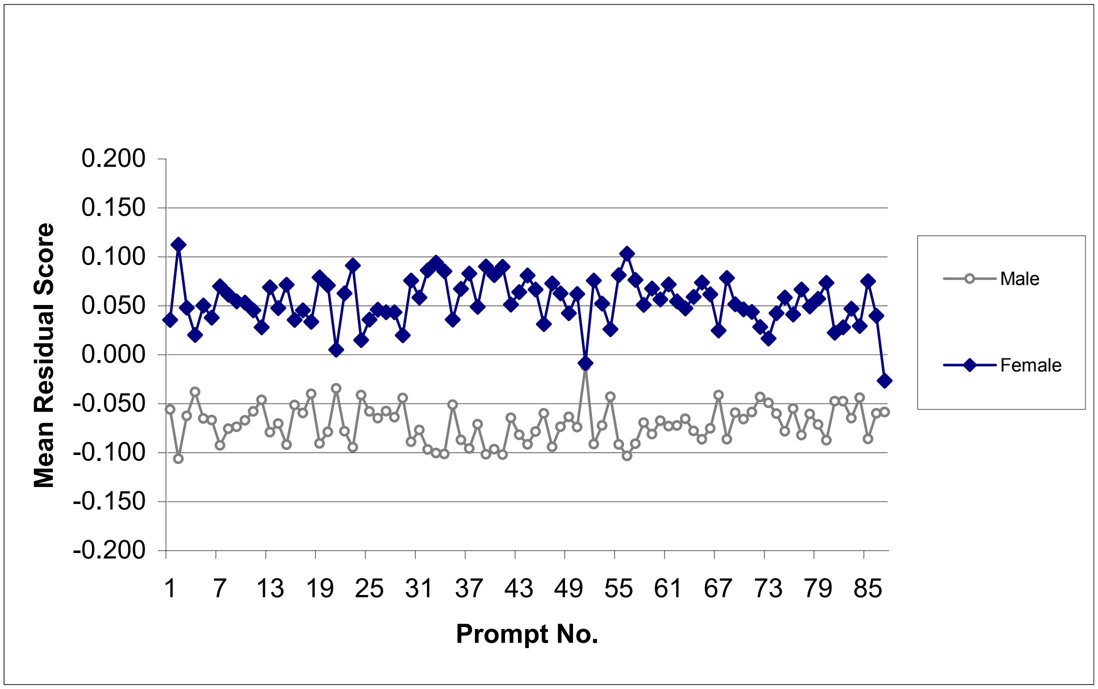

Figure 3. **Mean residual scores of male and female examinee groups after controlling for English language ability for each of the Phase II prompts.**

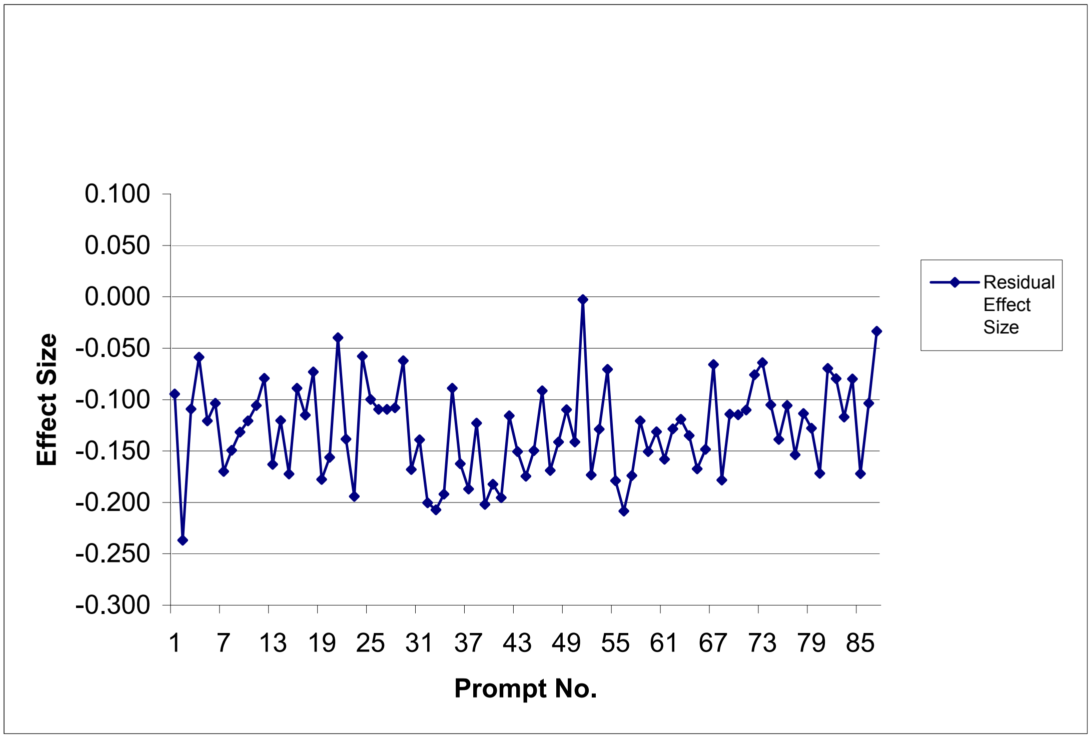

Figure 4. **Residual-based effect sizes for after controlling for English language ability for each of the Phase II prompts.**

Analyses of uniform and nonuniform effects. Table 2 gives the results of Zumbo's (1999) three-step modeling process and shows that one prompt had no gender effect, 17 prompts exhibited a significant ability-by-group interaction ($x* D_m$) indicating nonuniform gender effects, and 69 prompts displayed uniform gender effects only (see also Table D1).

Table 2

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

Group effect	# of prompts	English Language ability (x)		Gender group (D_m)		Ability x group interaction ($x* D_m$)	
		Mean β_1	Mean R^2	Mean β_2	Mean R^2	Mean β_3	Mean R^2
No effect	1	–0.53**	0.3880				
Uniform only	69	–0.52**	0.3727	–0.29*	0.3772		
Uniform + nonuniform							
Uniform dominant	17	–0.58**	0.3672	–0.32*	0.3728	0.04*	0.3735
Nonuniform dominant	0						
Total	87	–0.53**	0.3716				

Note. Based on Phase II analyses of 87 prompts.
* $p < 0.05$ two-tailed. ** $p < 0.01$ two-tailed.

Table 3 gives results for five prompts with the largest uniform gender effect sizes. Three of the prompts in Table 3 (2, 39, and 56) had no nonuniform gender effects, but the remaining two prompts (32 and 33) had both uniform and nonuniform gender effects. Prompt 2 had a total R^2 effect size of 0.0146, because there was no nonuniform effect. This gender effect is quite small by either Zumbo’s standard of .13 for a “negligible” effect or by Cohen’s standard of .02 for a “small” effect.

Table 4 gives the results for five prompts with the largest nonuniform gender effect sizes. The total gender effect sizes in Table 4 are all “negligible” by Zumbo’s standard and “small” by Cohen’s.

Table 3

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

Prompt no.	No. of examinees		Slope for gender group (β_2)	R-squared effect size		
	Male	Female		Uniform	Nonuniform	Total
Prompt 2	1,388	1,203	–0.52*	0.0146		0.0146
Prompt 32	3,407	2,962	–0.43*	0.0099	0.0010	0.0109
Prompt 33	5,429	4,617	–0.46*	0.0107	0.0006	0.0113
Prompt 39	3,457	3,025	–0.46*	0.0107		0.0107
Prompt 56	3,766	3,341	–0.47*	0.0106		0.0106

Note. Based on Phase II analyses of 87 prompts.
* p < 0.01 two-tailed.

Table 4

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

Prompt no.	No. of examinees		Slope for interaction term (β_2)	R-squared effect size		
	Male	Female		Uniform	Nonuniform	Total
Prompt 32	3,407	2,962	0.05*	0.0099	0.001	0.0109
Prompt 43	4,158	3,688	0.05*	0.006	0.0008	0.0068
Prompt 54	3,382	3,094	0.05*	0.0014	0.0009	0.0023
Prompt 63	4,227	3,791	0.05*	0.0038	0.0009	0.0047
Prompt 68	3,826	3,382	0.05*	0.0085	0.0008	0.0093

Note. Based on Phase II analyses of 87 prompts.
* p < 0.01 two-tailed.

Figures 5 and 6 present graphic representations of uniform and nonuniform effects. Figure 5 shows the uniform effect for Prompt 2 and Figure 6 the nonuniform effect for Prompt 32. Although the total R^2 gender effect size for Prompt 32 is only .0109 (as indicated in Table 4), a large proportion of this is uniform (.0099). The result is that the regression lines cross at an ELA level of 5, with the female group receiving higher expected scores in the low ranges of ELA and the male group receiving higher expected scores in the higher ranges of ELA. Figure 6 shows that, although the total gender effect for Prompt 32 is "negligible" or "small" in Zumbo's and Cohen's terminology, the expected score difference between gender groups is substantial at lower ELA score levels. For example at an ELA score level of –9.2, the gender effect size is about .41 standard deviations, "medium" effect size by Cohen's standard.

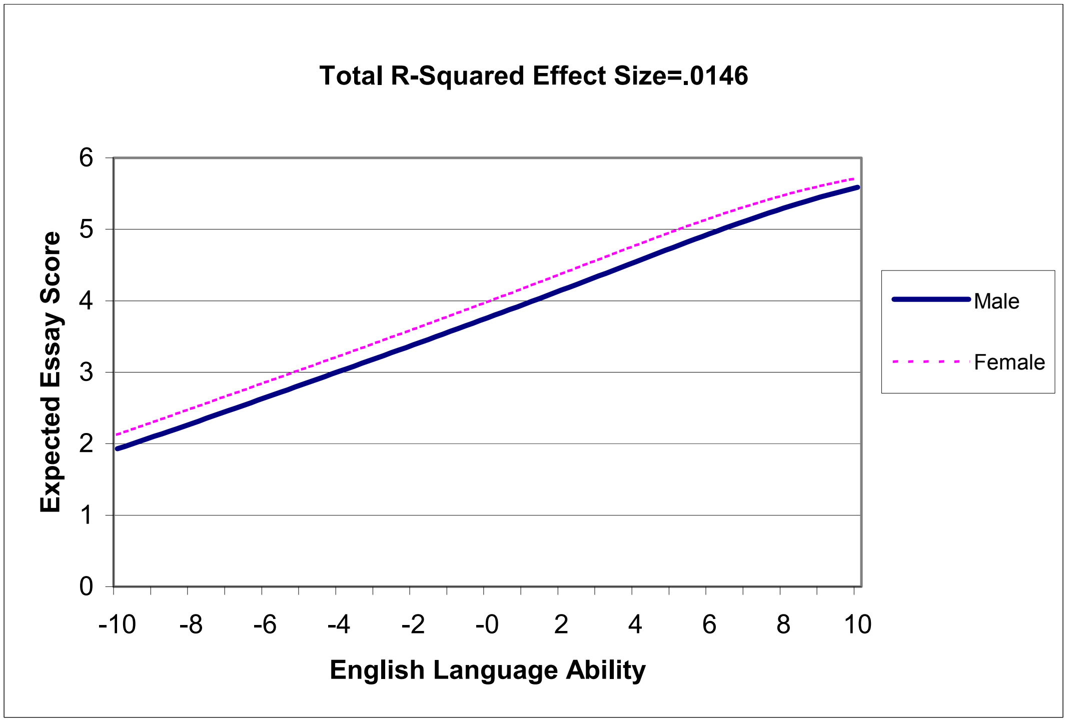

Figure 5. **Separate expected score curves for the reference and focal groups based on the full logistic regression model: Largest uniform effect (Prompt 2).**

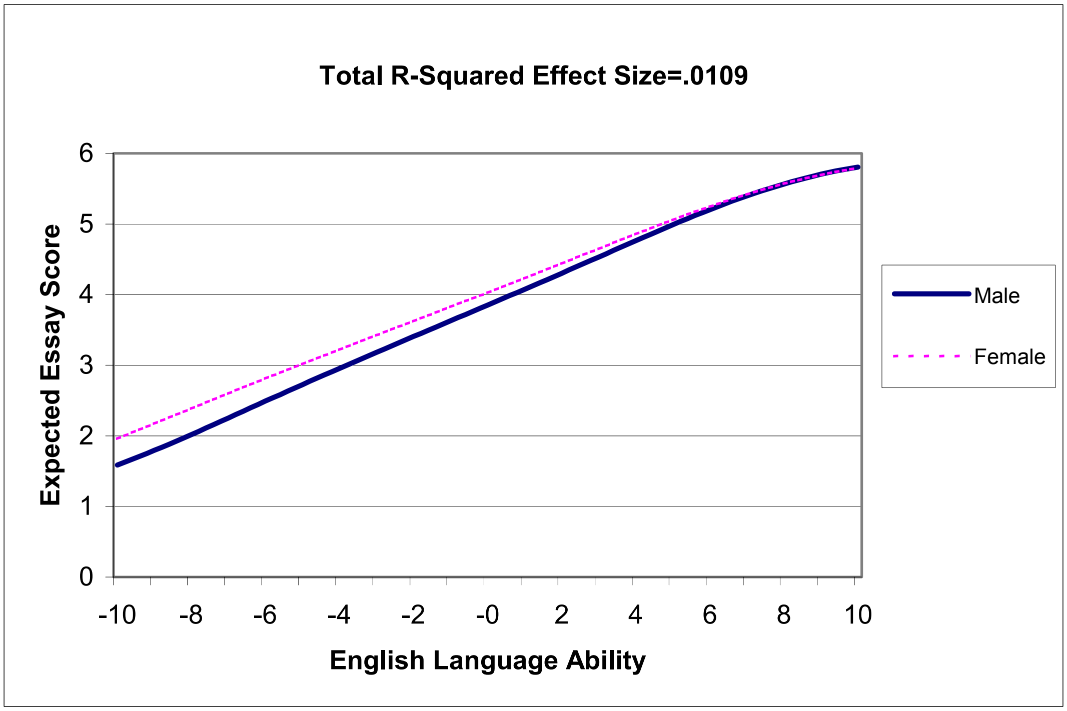

Figure 6. **Separate expected score curves for the reference and focal groups based on the full logistic regression model: Largest nonuniform effect (Prompt 32).**

More detailed results for all prompts are presented in Appendixes B, C, and D.

Prompt difficulty analyses. In addition to the analyses of gender differences reported above, analyses were also conducted to identify prompts that appeared to be most and least difficult for examinees. Table 5 presents a summary of the prompt difficulties for the 10 most difficult prompts and the 10 least difficult prompts. Difficulty indices for dichotomized polytomous responses (ξ_{jk}) and essay prompt ($\bar{\xi}_j$) were computed using the intercept and slope parameters from the Step 1 model.

Table 5

The 10 Most Difficult and Least Difficult TOEFL Writing Prompts Identified From the Logistic Regression Analyses (Step 1 Model)

Prompt no.	Difficulty category	*N*	Prompt difficulty		Effect size
			$\bar{\xi}$	$1/\bar{\xi}$	
29	Most	4,684	1.82	0.55	0.00
78		6,369	1.97	0.51	–0.03
79		6,674	1.99	0.50	0.06
81		6,843	2.01	0.50	0.07
22		6,453	2.05	0.49	0.08
10		7,217	2.11	0.47	0.04
41		8,825	2.12	0.47	0.04
56		7,107	2.26	0.44	0.10
42		8,131	2.26	0.44	0.12
28		5,424	2.27	0.44	0.12
32	Least	6,369	3.05	0.33	0.17
27		8,326	3.05	0.33	0.21
77		7,943	3.07	0.33	0.24
9		4,059	3.11	0.32	0.25
66		8,633	3.16	0.32	0.23
13		6,164	3.21	0.31	0.24
2		2,591	3.23	0.31	0.19
76		6,321	3.36	0.30	0.23
8		4,647	3.39	0.29	0.27
17		2,066	3.43	0.29	0.30

Note. Based on Phase II analyses of 87 prompts. Positive effect sizes in Table 5 indicate higher scores on the less difficult prompts.

Table 5 shows that prompt difficulty indices ranged from 3.43 (least difficult) to 1.82 (most difficult). An alternative statistic for prompt difficulty is the reciprocal of difficulty, which has a range for the 87 prompts examined from .55 (most difficult) to .29 (least difficult).

The effect size was computed as the mean difference in standard deviation units between the most difficult prompt (#29) and the other prompts. For example, the prompt indicated in Table 5 as being the least difficult (#17) has a mean score .29 standard deviations higher than prompt 29. Accordingly, the difference in difficulty between the most difficult and the least difficult TOEFL writing prompts examined was only slightly greater than the .20 magnitude that Cohen (1988, 1992) considered to be a “small” effect.

More detailed results of the prompt difficulty analysis are given in Appendix E.

Summary and Discussion

The comparability of TOEFL writing prompts was examined using a number of different methods. Mean essay and ELA score differences were examined for 47 different prompts in Phase I and 87 prompts in Phase II, including the prompts analyzed in Phase I. In Phase I, logistic regression analyses, controlling for examinee English language ability, were conducted to develop an index of prompt difficulty and to analyze gender differences for different prompts. An expert panel of applied linguists and TOEFL program staff reviewed prompts selected for extreme difficulties and gender differences. A taxonomy of TOEFL writing prompt characteristics was developed in Phase I using judgments made in the expert panel review and other information obtained from the data analyses.

In Phase I, TOEFL essay score distributions were plotted for all prompts examined, and skewness and kurtosis were computed for each. Four prompts with unusual distributions were selected for further analysis. The prompts selected were compared for central tendency and were found to be significantly different statistically. One prompt had a relatively high percentage of “4” scores, while another had a relatively low percentage of “4” scores. Two other prompts differed significantly in the number of scores in the 3.5 to 4.5 range. Effect sizes were all less than .20 standard deviations. Experts who reviewed these prompts offered no clear suggestions for the distributional differences, and no hypotheses were developed. Accordingly, the distributional differences were not considered to be of major importance.

The Phase II analyses showed that the difference between the highest mean score and the lowest mean score was only .30 standard deviations. Such a difference is generally considered to be a relatively small difference. Only nine prompts had mean score differences from other prompts exceeding .20 standard deviations. These prompt mean differences were not the result of English language ability differences for the prompt assignment groups, however, since the English language differences were quite small. The small differences in English language ability for the prompt assignment groups indicate that the randomization of prompt assignments was very effective.

Mean TOEFL essay score gender differences were also generally relatively small. The largest effect size was .24 standard deviations (favoring females), and only 3 of 87 prompts had effect sizes greater than .20. For the prompt with the largest gender effect size of .24 standard deviations, the mean English language ability difference for the two genders was only .02 standard deviations (favoring females). These findings of gender differences for essay tasks are consistent with previous research showing that females tend to perform better than males on such tasks, on average.

The logistic regression analyses controlled for examinee English language ability and identified a number of prompts at extremes of difficulty and gender differences. Because mean English language ability differences across prompts were not large, these analyses tended to identify many of the same prompts identified by the mean difference analyses in which English ability was not controlled, although the precise ordering of prompts differed. The prompts identified by the logistic regression analyses as having extreme gender differences tended also to be the same prompts identified in the mean difference analyses, with some differences in the ordering. Because of these differences in ordering, the more rigorous logistic regression methodology was used to select specific prompts for further review. Almost all prompts analyzed had statistically significant gender differences, but effect sizes were relatively small. In general, it was believed that the logistic regression procedures worked well in this project, although there has been some controversy concerning utility. For example, French and Miller (1996) considered the methodology to be unwieldy, awkward, time consuming, and cumbersome, and concluded that it was "difficult to recommend logistic regression for use as an omnibus DIF detection procedure for polytomous items."

Conclusions and Recommendations

Three conclusions or recommendations may be drawn from this study:

1. Although the differences are relatively small, the present pool of TOEFL CBT essay prompts contains prompts that are of statistically significant differential difficulty and that generate statistically significant gender differences. Although the differences observed are not large by accepted statistical standards, a policy should be formulated for what levels of difference should result in prompts being dropped from active administration.
2. Expert review of prompts at the extremes of difficulty and gender differences resulted in general agreement about what tends to characterize such prompts, but such characterizations did not always explain difficulty and gender differences.
3. Although expert review of prompts can indicate why some prompts may be less comparable than others, it is a relatively inefficient procedure and does not always explain why differences occur. It is therefore recommended that statistical quality control procedures be routinely implemented to identify less comparable prompts. These procedures need to be based on a defensible methodology coupled with a sound program policy. Prompt developers can benefit from the routine identification and review of extreme prompts identified through statistical quality control.

References

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

Bermudez, A. B., & Prater, D. L. (1994). Examining the effects of gender and second language proficiency on Hispanic writers' persuasive discourse. *Bilingual Research Journal, 18*(3–4), 47–62.

Berninger, V. W., & Fuller, F. (1992). Gender differences in orthographic, verbal, and compositional fluency—Implications for assessing writing disabilities in primary grade children. *Journal of School Psychology, 30*(4), 363–382.

Breland, H. M., & Griswold, P. A. (1982). Use of a performance test as a criterion in a differential validity study. *Journal of Educational Psychology, 74*(5), 713–721.

Breland, H. M., & Jenkins, L. M. (1997). *English word frequency statistics: Analysis of a corpus of 14 million tokens.* New York: College Entrance Examination Board.

Breland, H. M., & Jones, R. J. (1982). *Perceptions of writing skill* (College Board Report No. 88–3, ETS RR-82-47). New York: The College Board.

Bridgeman, B., & Bonner, M. (1994). *SAT-W as a predictor of grades in freshman English composition courses.* Unpublished manuscript.

Bridgeman, B., & McHale, F. (1996). Potential impact of the addition of writing assessment on admissions decisions. *Research in Higher Education, 39,* 663–677.

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

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98–101.

Cumming, A. (1990a). Expertise in evaluating second language composition. *Language Testing, 7*, 31–51.

Cumming, A. (1990b). Metalinguistic and ideational thinking in second language composing. *Written Communication, 7*, 482–511.

Cumming, A., & Mellow, D. (1996). An investigation into the validity of written indicators of second language proficiency. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 72–93). Clevedon, UK: Multilingual Matters.

Englehard, G., Gordon, B., & Gabrielson, S. (1991). The influences of mode of discourse, experiental demand, and gender on the quality of school writing. *Research in the Teaching of English, 26*(3), 315–336.

ETS. (1998). *Computer-based TOEFL test score user guide.* 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*(3), 315–332.

Golub-Smith, M., Reese, C., & Steinhaus, K. (1993). *Topic and topic type comparability on the Test of Written English* (TOEFL Research Report No. 42, ETS RR-93-10). Princeton, NJ: ETS.

Heck, R. H., & Crislip, M. (2001). Direct and indirect writing assessments: Examining issues of equity and utility. *Educational Evaluation and Policy Analysis, 23*(1), 19–36.

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

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.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). *Sex-related performance differences on constructed response and multiple-choice sections of Advanced Placement Examinations* (College Board Report No. 92-7, ETS RR-93-5). New York: The College Board.

National Assessment of Educational Progress. (1994). *NAEP 1992 trends in academic progress.* Washington, DC: National Center for Education Statistics.

Pomplun, M., Wright, D., Oleka, N., & Sudlow, M. (1992). *An analysis of English Composition Test prompts for differential difficulty* (College Board Report No. 92-4, ETS RR-92-34). New York: The College Board.

Powers, D. E., & Fowles, M. E. (1999). Test-takers' judgments of essay prompts: Perceptions and performance. *Educational Assessment, 6*(1), 3–22.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph, 17.*

SAS Institute. (1990). *SAS/STAT user's guide, version 6* (4th ed.). Cary, NC: SAS Institute, Inc.

Schaeffer, G. A., Briel, J. B., & Fowles, M. E. (2001). *Psychometric evaluation of the new GRE Writing Assessment* (GRE Board Profession Report No. 96–11P, ETS RR-01-08). Princeton, NJ: ETS.

Sincoff, J. B., & Sternberg, R. J. (1987). Two faces of verbal ability. *Intelligence, 11,* 263–276.

Sincoff, J. B., & Sternberg, R. J. (1988). Development of verbal fluency abilities and strategies in elementary-school-age children. *Developmental Psychology, 24,* 646–653.

Stansfield, C. W., & Ross, J. (1988). A long-term research agenda for the Test of Written English. *Language Testing, 5,* 160–186.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Erlbaum.

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.

Notes

[1] In an adaptive test, the computer determines which question is presented next based on examinee performance on preceding questions, whereas, in a linear test, questions are chosen without consideration of examinee performance on the previous questions (ETS, 1998).

Appendix A

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

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

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

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 gender covariate, as in our TOEFL essay data, we can rewrite the logit g_i (x,D) as:

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

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 gender 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 male examinee group and 1 for the female examinee group, respectively. It should be noted that the score-by-group interaction term was also added to examine the score difference of nonuniform nature between the two groups.

The dichotomous 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 (K-1). The proportional log-odds for each dichotomized response based on the cumulative logit scheme can be expressed as:

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

where L_{ij} stands for the proportional log-odds ratio for a dichotomized response (i) on the polytomous item (j), and k is a subscript of the response category (k=0,1,2,…,K) for an examinee score (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 (1969) graded response model.

Category Characteristic Curves and Prompt Difficulty Index

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

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

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 (Samejima, 1969), $P^+_{j0}(x,D)$ and $P^+_{j,K+1}(x,D)$ can also be 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

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

The essay prompt difficulty index can be derived from the simple model without the group variable as in equation 6, which is similar to the item location parameter in item response theory (IRT). The equation above can be rewritten as:

$$g_i(x) = \beta_{0i} + \beta_1 x = \beta_1(x + \frac{\beta_{0i}}{\beta_1}) = \beta_1(x - \xi_i) \qquad (6)$$

where ξ_i is analogous to an item category difficulty in IRT. Therefore, the mean of the ξ_i over all dichotomized responses can be interpreted as the item location parameter for the polytomous item and can be written as:

$$\overline{\xi_j} = \frac{1}{K-1}\sum_{i=1}^{K-1}\xi_{ji} \qquad (7)$$

Appendix B

Number of Essays, Means, and Standard Deviations of English Language Ability and Essay Scores for Male and Female Examinee Groups

Table B1

Number of Examinees for Male and Female Examinee Groups for 87 Prompts (Phase II)

Prompt no.	Phase I			Phase II		
	Male	Female	Total	Male	Female	Total
1	*	*	*	2,991	2,763	5,754
2	*	*	*	1,388	1,203	2,591
3	1,627	1,334	2,961	4,466	3,867	8,333
4	1,608	1,362	2,970	4,333	3,686	8,019
5	1,533	1,403	2,936	5,183	4,685	9,868
6	1,502	1,303	2,805	4,036	3,605	7,641
7	*	*	*	4,180	3,886	8,066
8	*	*	*	2,420	2,227	4,647
9	1,409	1,167	2,576	2,162	1,897	4,059
10	1,362	1,145	2,507	3,856	3,361	7,217
11	*	*	*	4,168	3,910	8,078
12	*	*	*	2,818	2,652	5,470
13	*	*	*	3,249	2,915	6,164
14	*	*	*	3,300	3,004	6,304
15	1,392	1,263	2,655	5,401	4,911	10,312
16	*	*	*	1,269	1,136	2,405
17	*	*	*	1,041	1,025	2,066
18	*	*	*	2,477	2,262	4,739
19	*	*	*	3,183	2,822	6,005
20	*	*	*	1,208	1,117	2,325
21	1,602	1,369	2,971	3,967	3,532	7,499
22	*	*	*	3,443	3,010	6,453
23	1,601	1,325	2,926	4,449	3,847	8,296
24	*	*	*	4,047	3,645	7,692
25	*	*	*	3,931	3,476	7,407
26	*	*	*	3,045	2,756	5,801
27	1,319	1,081	2,400	4,492	3,834	8,326
28	*	*	*	2,878	2,546	5,424
29	1,354	1,084	2,438	2,539	2,145	4,684

(Table continues)

Table B1 (continued)

Prompt	Phase I			Phase II		
no.	Male	Female	Total	Male	Female	Total
30	1,596	1,297	2,893	3,965	3,333	7,298
31	1,249	1,044	2,293	3,719	3,232	6,951
32	*	*	*	3,407	2,962	6,369
33	1,492	1,261	2,753	5,429	4,617	10,046
34	1,424	1,210	2,634	4,692	4,087	8,779
35	1,235	1,083	2,318	4,043	3,660	7,703
36	*	*	*	3,112	2,673	5,785
37	1,254	1,139	2,393	3,681	3,204	6,885
38	1,388	1,187	2,575	5,011	4,510	9,521
39	1,108	939	2,047	3,457	3,025	6,482
40	1,333	1,111	2,444	3,868	3,312	7,180
41	1,462	1,265	2,727	4,715	4,110	8,825
42	1,068	967	2,035	4,344	3,787	8,131
43	*	*	*	4,158	3,688	7,846
44	*	*	*	4,976	4,415	9,391
45	*	*	*	4,048	3,581	7,629
46	2,166	2,120	4,286	4,889	4,501	9,390
47	2,140	1,931	4,071	6,199	5,561	11,760
48	*	*	*	2,998	2,693	5,691
49	1,973	1,814	3,787	5,764	5,211	10,975
50	1,748	1,481	3,229	5,524	4,858	10,382
51	1,967	1,718	3,685	4,103	3,668	7,771
52	2,143	2,044	4,187	4,353	3,926	8,279
53	*	*	*	3,480	3,161	6,641
54	*	*	*	3,382	3,094	6,476
55	*	*	*	3,989	3,385	7,374
56	*	*	*	3,766	3,341	7,107
57	1,890	1,591	3,481	4,974	4,219	9,193
58	1,963	1,793	3,756	4,556	4,119	8,675
59	2,086	1,808	3,894	5,753	4,941	10,694
60	*	*	*	3,438	2,820	6,258
61	1,849	1,542	3,391	4,894	4,149	9,043
62	*	*	*	3,035	2,568	5,603
63	*	*	*	4,227	3,791	8,018
64	2,116	1,860	3,976	3,582	3,116	6,698
65	1,700	1,488	3,188	4,616	4,083	8,699

(Table continues)

Table B1 (continued)

Prompt	Phase I			Phase II		
no.	Male	Female	Total	Male	Female	Total
66	1,458	1,289	2,747	4,526	4,107	8,633
67	2,034	1,779	3,813	4,592	4,054	8,646
68	1,636	1,442	3,078	3,826	3,382	7,208
69	1,699	1,410	3,109	4,374	3,750	8,124
70	*	*	*	4,273	3,895	8,168
71	*	*	*	4,317	3,796	8,113
72	1,420	1,178	2,598	5,923	5,088	11,011
73	1,774	1,511	3,285	4,001	3,425	7,426
74	1,542	1,294	2,836	3,967	3,363	7,330
75	1,728	1,463	3,191	4,487	3,829	8,316
76	*	*	*	3,397	2,924	6,321
77	1,683	1,400	3,083	4,273	3,670	7,943
78	*	*	*	3,330	3,039	6,369
79	1,470	1,184	2,654	3,629	3,045	6,674
80	*	*	*	2,781	2,452	5,233
81	1,510	1,254	2,764	3,697	3,146	6,843
82	1,488	1,196	2,684	4,212	3,573	7,785
83	*	*	*	2,706	2,405	5,111
84	*	*	*	2,599	2,168	4,767
85	*	*	*	3,002	2,647	5,649
86	*	*	*	2,756	2,430	5,186
87	1,389	1,191	2,580	5,418	4,779	10,197
Total	75,490	65,120	140,610	336,153	296,093	632,246
Mean	1,606	1,386	2,992	3,864	3,403	7,267
SD	286	285	567	1,039	913	1,949

Note. Asterisks (*) in some of the cells mean that no examinee data for the particular prompt were available in the Phase I study.

Table B2

Mean English Language Ability Scores for the Male and Female Examinee Groups for 87 Prompts (Phase II)

Prompt no.	Phase I				Phase II			
	Male		Female		Male		Female	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	*	*	*	*	0.09	2.79	0.05	2.64
2	*	*	*	*	–0.01	2.69	0.09	2.52
3	0.04	2.78	0.04	2.60	–0.02	2.77	–0.01	2.65
4	0.03	2.83	0.08	2.63	0.10	2.80	0.15	2.61
5	0.01	2.78	–0.05	2.63	0.04	2.77	0.11	2.57
6	–0.06	2.87	–0.07	2.63	0.09	2.81	0.06	2.64
7	*	*	*	*	–0.16	2.79	–0.12	2.62
8	*	*	*	*	0.00	2.78	0.16	2.56
9	0.15	2.76	0.13	2.57	0.12	2.75	0.07	2.60
10	0.15	2.75	–0.07	2.55	0.03	2.78	0.00	2.58
11	*	*	*	*	–0.11	2.82	–0.02	2.60
12	*	*	*	*	0.05	2.79	–0.01	2.58
13	*	*	*	*	–0.04	2.84	–0.08	2.68
14	*	*	*	*	0.01	2.79	0.01	2.63
15	0.05	2.75	–0.11	2.69	–0.01	2.80	0.04	2.59
16	*	*	*	*	–0.12	2.81	–0.02	2.62
17	*	*	*	*	0.05	2.74	0.07	2.56
18	*	*	*	*	0.03	2.83	0.08	2.64
19	*	*	*	*	0.03	2.78	0.06	2.68
20	*	*	*	*	–0.07	2.72	–0.10	2.63
21	0.03	2.81	0.05	2.65	0.08	2.78	0.01	2.65
22	*	*	*	*	–0.06	2.85	0.03	2.70
23	0.07	2.76	0.12	2.59	0.02	2.80	0.12	2.57
24	*	*	*	*	–0.08	2.78	–0.03	2.61
25	*	*	*	*	–0.04	2.74	–0.02	2.62
26	*	*	*	*	–0.12	2.83	–0.01	2.64
27	0.05	2.79	0.08	2.57	0.05	2.80	0.11	2.57
28	*	*	*	*	0.04	2.75	0.13	2.53
29	–0.03	2.84	0.00	2.63	0.02	2.81	0.09	2.59
30	0.02	2.83	–0.05	2.63	0.11	2.83	0.07	2.58
31	0.15	2.72	–0.05	2.65	0.07	2.75	0.07	2.57
32	*	*	*	*	–0.19	2.81	–0.07	2.61
33	0.10	2.78	–0.04	2.64	–0.01	2.77	–0.02	2.61
34	0.14	2.72	0.07	2.66	0.06	2.74	0.11	2.60
35	0.17	2.78	–0.05	2.61	0.03	2.78	0.03	2.57

(Table continues)

Table B2 (continued)

Prompt	Phase I				Phase II			
no.	Male		Female		Male		Female	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
36	*	*	*	*	–0.01	2.78	0.00	2.60
37	0.04	2.84	0.10	2.57	0.06	2.78	0.06	2.62
38	0.21	2.79	0.14	2.53	0.05	2.80	0.10	2.58
39	0.27	2.75	0.03	2.66	0.17	2.76	0.14	2.61
40	0.15	2.67	0.09	2.61	0.17	2.76	0.17	2.57
41	–0.02	2.78	–0.05	2.71	0.03	2.77	0.05	2.59
42	0.07	2.78	0.09	2.72	0.07	2.79	0.17	2.60
43	*	*	*	*	–0.01	2.79	–0.03	2.61
44	*	*	*	*	–0.13	2.80	–0.11	2.68
45	*	*	*	*	–0.19	2.79	–0.10	2.62
46	0.01	2.83	–0.02	2.66	0.01	2.83	0.07	2.59
47	–0.02	2.81	0.00	2.65	0.07	2.78	0.07	2.66
48	*	*	*	*	–0.10	2.83	–0.11	2.61
49	0.03	2.75	–0.07	2.70	0.09	2.76	0.05	2.63
50	0.01	2.82	0.00	2.67	0.01	2.83	0.02	2.64
51	0.04	2.77	0.02	2.62	0.10	2.79	0.08	2.62
52	–0.01	2.78	–0.07	2.69	0.13	2.78	0.09	2.62
53	*	*	*	*	–0.03	2.75	0.13	2.55
54	*	*	*	*	0.02	2.68	0.14	2.53
55	*	*	*	*	–0.07	2.77	0.11	2.55
56	*	*	*	*	–0.05	2.82	–0.03	2.60
57	0.00	2.82	–0.03	2.69	0.03	2.81	0.02	2.64
58	–0.07	2.85	–0.06	2.58	0.01	2.80	0.02	2.61
59	0.01	2.80	–0.09	2.65	0.03	2.78	–0.02	2.62
60	*	*	*	*	–0.17	2.77	–0.08	2.61
61	0.04	2.86	0.02	2.65	0.02	2.82	0.03	2.61
62	*	*	*	*	–0.03	2.84	–0.08	2.62
63	*	*	*	*	–0.04	2.80	–0.02	2.61
64	0.06	2.73	0.00	2.63	0.13	2.73	0.02	2.65
65	–0.09	2.82	–0.02	2.70	0.05	2.77	0.09	2.61
66	0.19	2.73	–0.06	2.67	0.08	2.73	–0.01	2.64
67	–0.04	2.83	–0.11	2.60	0.07	2.79	0.03	2.59
68	0.06	2.77	0.03	2.63	0.06	2.76	0.07	2.60
69	0.06	2.82	–0.13	2.73	0.05	2.81	–0.01	2.66
70	*	*	*	*	0.10	2.77	0.03	2.60
71	*	*	*	*	–0.15	2.84	–0.04	2.57
72	0.19	2.82	0.01	2.66	0.03	2.82	0.07	2.63
73	0.08	2.72	–0.10	2.62	0.10	2.75	–0.02	2.64
74	0.01	2.72	–0.03	2.62	0.02	2.73	0.04	2.59
75	–0.08	2.79	–0.13	2.63	–0.05	2.76	–0.05	2.60

(Table continues)

Table B2 (continued)

Prompt no.	Phase I				Phase II			
	Male		Female		Male		Female	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
76	*	*	*	*	0.01	2.79	0.00	2.67
77	0.08	2.78	–0.01	2.67	0.10	2.80	0.05	2.61
78	*	*	*	*	–0.20	2.80	–0.08	2.59
79	0.27	2.72	–0.17	2.77	0.19	2.76	0.00	2.65
80	*	*	*	*	–0.06	2.77	–0.01	2.63
81	0.01	2.81	0.05	2.61	0.12	2.72	0.21	2.55
82	0.00	2.86	–0.02	2.56	–0.02	2.81	0.00	2.62
83	*	*	*	*	–0.09	2.82	0.02	2.62
84	*	*	*	*	0.05	2.76	0.00	2.65
85	*	*	*	*	–0.05	2.80	–0.03	2.66
86	*	*	*	*	–0.03	2.86	0.02	2.64
87	–0.07	2.84	0.02	2.62	–0.08	2.76	–0.03	2.59
Mean	0.05	2.79	–0.01	2.64	0.01	2.78	0.03	2.61

Note. Asterisks (*) in some of the cells mean that no examinee data for the particular prompt were available in the Phase I study.

Table B3

Mean Raw Essay Scores for the Male and Female Examinee Groups for 87 Prompts (Phase II)

Prompt no.	Phase I				Phase II			
	Male		Female		Male		Female	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	*	*	*	*	4.07	0.99	4.16	0.95
2	*	*	*	*	3.99	0.96	4.22	0.88
3	3.99	1.07	4.13	1.02	3.99	1.03	4.10	1.00
4	4.06	1.08	4.14	0.99	4.06	1.02	4.13	0.95
5	4.02	1.03	4.10	0.96	4.00	0.98	4.13	0.93
6	3.99	1.09	4.09	1.02	3.99	1.03	4.09	0.98
7	*	*	*	*	4.01	0.99	4.18	0.92
8	*	*	*	*	4.08	0.95	4.25	0.88
9	4.15	1.03	4.27	0.96	4.11	1.00	4.22	0.94
10	3.94	1.05	4.05	0.98	3.91	1.02	4.02	0.98
11	*	*	*	*	3.94	1.00	4.06	0.94
12	*	*	*	*	4.00	0.96	4.06	0.91
13	*	*	*	*	4.07	0.91	4.21	0.90
14	*	*	*	*	3.98	1.00	4.10	0.95
15	3.91	0.99	4.05	0.98	3.91	0.97	4.08	0.92

(Table continues)

Table B3 (continued)

Prompt	Phase I				Phase II			
no.	Male		Female		Male		Female	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
16	*	*	*	*	4.03	1.00	4.14	0.95
17	*	*	*	*	4.14	0.92	4.25	0.90
18	*	*	*	*	4.00	1.02	4.08	0.99
19	*	*	*	*	4.05	0.99	4.22	0.92
20	*	*	*	*	4.04	0.95	4.18	0.96
21	4.04	1.04	4.07	0.99	4.05	1.01	4.07	0.97
22	*	*	*	*	3.92	1.04	4.08	0.99
23	3.87	1.00	4.10	0.97	3.90	0.97	4.10	0.94
24	*	*	*	*	4.05	1.00	4.12	0.94
25	*	*	*	*	4.03	0.96	4.13	0.91
26	*	*	*	*	3.97	1.03	4.11	1.00
27	4.09	1.00	4.18	0.91	4.06	0.94	4.17	0.90
28	*	*	*	*	3.97	1.02	4.10	0.97
29	3.90	1.10	4.00	1.01	3.88	1.07	3.96	0.99
30	3.93	1.06	4.06	0.99	3.99	1.00	4.14	0.95
31	4.08	1.00	4.20	0.97	4.05	1.00	4.18	0.94
32	*	*	*	*	3.98	0.95	4.18	0.87
33	4.07	1.02	4.22	0.94	4.01	0.97	4.20	0.90
34	3.99	1.04	4.20	0.99	4.01	0.99	4.21	0.95
35	4.03	1.04	4.10	0.99	4.03	1.01	4.12	0.94
36	*	*	*	*	3.97	0.97	4.12	0.92
37	3.95	1.01	4.18	0.95	3.98	0.97	4.16	0.94
38	4.11	1.05	4.25	0.97	4.06	1.00	4.19	0.94
39	4.02	0.97	4.21	0.97	4.00	0.97	4.18	0.93
40	4.01	1.02	4.16	1.00	4.02	0.99	4.19	0.96
41	3.86	1.05	4.05	1.00	3.86	1.01	4.06	0.95
42	3.99	1.06	4.15	1.03	3.98	1.03	4.11	0.97
43	*	*	*	*	3.96	1.00	4.10	0.93
44	*	*	*	*	3.89	1.01	4.06	0.96
45	*	*	*	*	3.87	0.99	4.04	0.94
46	4.02	1.07	4.15	0.98	4.03	1.03	4.14	0.95
47	3.99	1.06	4.21	1.00	3.99	1.00	4.15	0.97
48	*	*	*	*	3.95	0.98	4.09	0.94
49	4.06	1.02	4.17	1.00	4.03	0.98	4.13	0.95
50	4.04	1.00	4.17	0.96	4.00	0.99	4.14	0.93
51	4.06	1.05	4.08	1.01	4.07	1.04	4.07	0.98
52	3.96	1.00	4.11	0.99	4.00	0.98	4.16	0.95
53	*	*	*	*	3.99	0.99	4.14	0.94
54	*	*	*	*	4.01	1.01	4.11	0.94

(Table continues)

Table B3 (continued)

Prompt no.	Phase I				Phase II			
	Male		Female		Male		Female	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
55	*	*	*	*	3.92	0.98	4.13	0.95
56	*	*	*	*	3.92	1.02	4.13	0.96
57	3.98	0.99	4.15	0.95	3.96	0.98	4.13	0.94
58	3.92	1.07	4.06	0.97	3.93	1.01	4.05	0.97
59	4.01	1.05	4.15	1.01	4.02	1.01	4.15	0.96
60	*	*	*	*	3.98	0.96	4.12	0.92
61	3.97	0.97	4.17	0.93	4.00	0.94	4.14	0.89
62	*	*	*	*	4.02	1.02	4.13	0.95
63	*	*	*	*	3.99	0.97	4.10	0.91
64	4.00	1.06	4.12	1.02	3.98	1.02	4.09	1.01
65	3.94	1.03	4.16	0.96	4.01	0.98	4.17	0.93
66	4.15	0.97	4.20	0.91	4.08	0.94	4.20	0.90
67	3.94	1.07	3.98	1.01	3.98	1.02	4.03	0.98
68	3.98	0.97	4.17	0.93	3.98	0.95	4.15	0.90
69	4.02	1.02	4.16	1.00	4.03	0.99	4.12	0.95
70	*	*	*	*	4.00	1.00	4.10	0.95
71	*	*	*	*	3.94	0.94	4.07	0.92
72	4.07	1.02	4.06	0.93	3.98	0.96	4.06	0.92
73	4.01	1.07	4.08	1.04	4.03	1.04	4.07	1.00
74	3.93	1.02	4.05	0.99	3.96	0.99	4.07	0.96
75	3.92	1.06	4.03	0.99	3.92	1.01	4.05	0.96
76	*	*	*	*	4.08	0.92	4.18	0.90
77	4.04	1.05	4.20	0.98	4.08	1.00	4.22	0.93
78	*	*	*	*	3.83	0.98	3.96	0.95
79	3.92	1.07	3.99	1.01	3.93	1.03	4.02	0.98
80	*	*	*	*	3.92	0.95	4.09	0.92
81	3.92	1.07	4.02	1.00	3.95	1.03	4.04	0.97
82	4.10	1.02	4.14	0.93	4.07	0.97	4.15	0.91
83	*	*	*	*	3.93	0.97	4.06	0.94
84	*	*	*	*	4.02	0.93	4.08	0.91
85	*	*	*	*	3.96	0.95	4.13	0.92
86	*	*	*	*	3.98	0.98	4.09	0.95
87	3.98	1.05	4.10	0.97	3.94	0.99	4.04	0.91
Mean	4.00	1.04	4.12	0.98	3.99	0.99	4.12	0.94

Note. Asterisks (*) in some of the cells mean that no examinee data for the particular prompt were available in the Phase I study.

Appendix C

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

Table C1 shows the mean expected essay scores, residuals, and standardized mean group differences after controlling for English language ability differences using the logistic regression step 1 model:

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

Table C1

Mean Expected Essay Scores and Residual–based Effect Sizes for 87 Prompts (Phase II)

Prompt	Expected essay scores				Residual (observed-expected)				Mean	Pooled	Residual
no.	Male		Female		Male		Female		resid.	SD	effect
	M	SD	M	SD	M	SD	M	SD	diff.	(obs)	size
1	4.13	0.60	4.12	0.57	–0.06	0.78	0.04	0.75	–0.09	0.97	–0.09
2	4.09	0.52	4.11	0.49	–0.11	0.79	0.11	0.75	–0.22	0.92	–0.24
3	4.05	0.64	4.05	0.61	–0.06	0.80	0.05	0.80	–0.11	1.01	–0.11
4	4.10	0.67	4.11	0.62	–0.04	0.76	0.02	0.73	–0.06	0.99	–0.06
5	4.06	0.60	4.08	0.56	–0.07	0.77	0.05	0.74	–0.12	0.96	–0.12
6	4.06	0.65	4.05	0.61	–0.07	0.80	0.04	0.77	–0.10	1.01	–0.10
7	4.10	0.61	4.11	0.58	–0.09	0.76	0.07	0.73	–0.16	0.96	–0.17
8	4.16	0.56	4.19	0.51	–0.08	0.76	0.06	0.72	–0.14	0.92	–0.15
9	4.18	0.61	4.17	0.58	–0.07	0.78	0.05	0.76	–0.13	0.98	–0.13
10	3.97	0.64	3.96	0.59	–0.07	0.79	0.05	0.77	–0.12	1.00	–0.12
11	3.99	0.63	4.01	0.58	–0.06	0.77	0.05	0.75	–0.10	0.98	–0.11
12	4.05	0.58	4.03	0.54	–0.05	0.75	0.03	0.74	–0.07	0.94	–0.08
13	4.15	0.57	4.14	0.54	–0.08	0.72	0.07	0.71	–0.15	0.91	–0.16
14	4.05	0.64	4.05	0.60	–0.07	0.77	0.05	0.74	–0.12	0.98	–0.12
15	4.00	0.59	4.01	0.54	–0.09	0.76	0.07	0.75	–0.16	0.95	–0.17

(Table continues)

Table C1 (continued)

Prompt	Expected essay scores				Residual (observed-expected)				Mean	Pooled	Residual
no.	Male		Female		Male		Female		resid.	SD	effect
	M	SD	M	SD	M	SD	M	SD	diff.	(Obs)	size
16	4.08	0.63	4.10	0.59	–0.05	0.77	0.04	0.75	–0.09	0.98	–0.09
17	4.20	0.57	4.20	0.54	–0.06	0.72	0.05	0.72	–0.10	0.91	–0.11
18	4.04	0.67	4.05	0.63	–0.04	0.76	0.03	0.77	–0.07	1.01	–0.07
19	4.14	0.59	4.14	0.56	–0.09	0.77	0.08	0.74	–0.17	0.95	–0.18
20	4.12	0.59	4.11	0.57	–0.08	0.76	0.07	0.76	–0.15	0.96	–0.16
21	4.08	0.64	4.06	0.61	–0.03	0.77	0.01	0.76	–0.04	0.99	–0.04
22	4.00	0.71	4.02	0.67	–0.08	0.75	0.06	0.75	–0.14	1.02	–0.14
23	3.99	0.60	4.01	0.55	–0.09	0.76	0.09	0.76	–0.19	0.96	–0.19
24	4.09	0.60	4.10	0.56	–0.04	0.79	0.02	0.76	–0.06	0.97	–0.06
25	4.09	0.58	4.09	0.55	–0.06	0.76	0.04	0.73	–0.09	0.94	–0.10
26	4.03	0.69	4.06	0.65	–0.06	0.76	0.05	0.76	–0.11	1.01	–0.11
27	4.12	0.57	4.13	0.52	–0.06	0.75	0.04	0.73	–0.10	0.92	–0.11
28	4.04	0.65	4.06	0.60	–0.06	0.77	0.04	0.77	–0.11	1.00	–0.11
29	3.92	0.66	3.94	0.61	–0.04	0.81	0.02	0.79	–0.06	1.03	–0.06
30	4.08	0.62	4.07	0.57	–0.09	0.78	0.08	0.76	–0.16	0.98	–0.17
31	4.12	0.61	4.12	0.57	–0.08	0.77	0.06	0.77	–0.14	0.97	–0.14
32	4.08	0.55	4.10	0.51	–0.10	0.75	0.09	0.73	–0.18	0.91	–0.20
33	4.11	0.57	4.11	0.54	–0.10	0.77	0.09	0.73	–0.19	0.94	–0.21
34	4.12	0.60	4.13	0.57	–0.10	0.78	0.09	0.77	–0.19	0.97	–0.19
35	4.08	0.64	4.08	0.60	–0.05	0.77	0.04	0.74	–0.09	0.98	–0.09
36	4.06	0.57	4.06	0.54	–0.09	0.77	0.07	0.75	–0.15	0.95	–0.16
37	4.07	0.59	4.08	0.56	–0.10	0.76	0.08	0.76	–0.18	0.96	–0.19
38	4.13	0.62	4.14	0.58	–0.07	0.78	0.05	0.75	–0.12	0.98	–0.12
39	4.10	0.58	4.09	0.55	–0.10	0.77	0.09	0.75	–0.19	0.95	–0.20
40	4.11	0.61	4.11	0.57	–0.10	0.77	0.08	0.77	–0.18	0.98	–0.18
41	3.96	0.62	3.97	0.58	–0.10	0.79	0.09	0.76	–0.19	0.98	–0.20
42	4.04	0.66	4.12	0.57	–0.06	0.78	0.05	0.75	–0.12	1.00	–0.12
43	4.04	0.61	4.11	0.49	–0.08	0.77	0.06	0.75	–0.15	0.97	–0.15
44	3.98	0.60	4.05	0.61	–0.09	0.80	0.08	0.78	–0.17	0.99	–0.17
45	3.95	0.61	4.06	0.62	–0.08	0.77	0.07	0.76	–0.14	0.97	–0.15
46	4.09	0.63	4.04	0.57	–0.06	0.80	0.03	0.77	–0.09	1.00	–0.09
47	4.08	0.61	3.98	0.57	–0.09	0.79	0.07	0.77	–0.17	0.99	–0.17
48	4.02	0.61	3.97	0.58	–0.07	0.76	0.06	0.76	–0.14	0.96	–0.14
49	4.10	0.61	4.10	0.58	–0.06	0.75	0.04	0.75	–0.11	0.96	–0.11
50	4.08	0.61	4.08	0.58	–0.07	0.77	0.06	0.75	–0.14	0.96	–0.14

(Table continues)

Table C1 (continued)

Prompt	Expected essay scores				Residual (observed-expected)				Mean	Pooled	Residual
no.	Male		Female		Male		Female		resid.	SD	effect
	M	SD	M	SD	M	SD	M	SD	diff.	(obs)	size
51	4.08	0.65	4.02	0.56	–0.01	0.79	–0.01	0.78	0.00	1.01	0.00
52	4.09	0.61	4.09	0.59	–0.09	0.76	0.08	0.74	–0.17	0.96	–0.17
53	4.06	0.60	4.08	0.56	–0.07	0.78	0.05	0.76	–0.12	0.97	–0.13
54	4.05	0.63	4.07	0.61	–0.04	0.77	0.03	0.75	–0.07	0.98	–0.07
55	4.01	0.60	4.08	0.58	–0.09	0.77	0.08	0.77	–0.17	0.97	–0.18
56	4.02	0.65	4.09	0.55	–0.10	0.77	0.10	0.76	–0.21	0.99	–0.21
57	4.05	0.61	4.08	0.59	–0.09	0.76	0.08	0.75	–0.17	0.96	–0.17
58	4.00	0.64	4.05	0.56	–0.07	0.78	0.05	0.78	–0.12	1.00	–0.12
59	4.10	0.62	4.03	0.60	–0.08	0.79	0.07	0.76	–0.15	0.99	–0.15
60	4.05	0.57	4.05	0.57	–0.07	0.76	0.06	0.75	–0.12	0.94	–0.13
61	4.07	0.55	4.00	0.59	–0.07	0.74	0.07	0.74	–0.14	0.92	–0.16
62	4.09	0.64	4.09	0.58	–0.07	0.78	0.05	0.75	–0.13	0.99	–0.13
63	4.05	0.60	4.07	0.54	–0.07	0.75	0.05	0.74	–0.11	0.95	–0.12
64	4.05	0.61	4.07	0.51	–0.08	0.81	0.06	0.81	–0.14	1.01	–0.13
65	4.09	0.58	4.08	0.59	–0.09	0.78	0.07	0.74	–0.16	0.96	–0.17
66	4.15	0.55	4.06	0.56	–0.08	0.75	0.06	0.73	–0.14	0.92	–0.15
67	4.02	0.64	4.03	0.59	–0.04	0.79	0.02	0.78	–0.07	1.00	–0.07
68	4.07	0.56	4.10	0.55	–0.09	0.74	0.08	0.74	–0.16	0.92	–0.18
69	4.09	0.61	4.13	0.53	–0.06	0.77	0.05	0.75	–0.11	0.97	–0.11
70	4.07	0.63	4.01	0.60	–0.07	0.77	0.05	0.75	–0.11	0.98	–0.11
71	4.00	0.56	4.07	0.53	–0.06	0.75	0.04	0.75	–0.10	0.93	–0.11
72	4.03	0.57	4.07	0.57	–0.04	0.77	0.03	0.75	–0.07	0.94	–0.08
73	4.08	0.63	4.05	0.59	–0.05	0.81	0.02	0.80	–0.07	1.02	–0.06
74	4.02	0.61	4.02	0.51	–0.06	0.77	0.04	0.76	–0.10	0.98	–0.11
75	4.00	0.62	4.03	0.53	–0.08	0.79	0.06	0.75	–0.14	0.99	–0.14
76	4.14	0.56	4.05	0.60	–0.05	0.72	0.04	0.72	–0.10	0.91	–0.11
77	4.17	0.60	4.03	0.58	–0.08	0.78	0.07	0.75	–0.15	0.97	–0.15
78	3.89	0.62	3.99	0.59	–0.06	0.76	0.05	0.75	–0.11	0.97	–0.11
79	4.01	0.66	4.14	0.53	–0.07	0.77	0.06	0.76	–0.13	1.01	–0.13
80	4.00	0.58	4.16	0.56	–0.09	0.75	0.07	0.74	–0.16	0.94	–0.17
81	3.99	0.65	3.91	0.58	–0.05	0.79	0.02	0.77	–0.07	1.01	–0.07
82	4.12	0.58	3.96	0.63	–0.05	0.76	0.03	0.74	–0.08	0.94	–0.08
83	3.99	0.61	4.01	0.55	–0.06	0.76	0.05	0.73	–0.11	0.95	–0.12
84	4.06	0.57	4.02	0.61	–0.04	0.72	0.03	0.73	–0.07	0.92	–0.08
85	4.05	0.58	4.12	0.55	–0.09	0.75	0.08	0.74	–0.16	0.94	–0.17
86	4.04	0.62	4.01	0.57	–0.06	0.75	0.04	0.75	–0.10	0.97	–0.10
87	3.99	0.58	4.05	0.55	–0.06	0.78	–0.03	2.59	–0.03	0.95	–0.03
Mean	4.06	0.61	4.05	0.55	–0.07	0.77	0.05	0.77	–0.12	0.97	–0.13

Appendix D

Uniform and Nonuniform Effect Sizes

Tables D1 and D2 show uniform and nonuniform effect sizes based on R^2 values for English language ability, gender group, and English-language-ability-by-gender-group interaction terms from the full (Step 3) logistic regression model:

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

Table D1

Uniform, Nonuniform, and Total R^2 Effect Sizes for 87 Prompts (Phase II)

Prompt	R^2 changes						$\chi 2$ test for added terms					
no.	R^2 values			R^2 effect size			Ability (A)		Group (G)		A*G	
	Ability	Group	A*G	Uni	Non	Total	$\chi 2$	*p*	$\chi 2$	*p*	$\chi 2$	*p*
1	0.3640	0.3665		0.0025		0.0025	2,034.79	< .0001	22.72	< .0001		
2	0.2981	0.3127		0.0146		0.0146	752.43	< .0001	53.77	< .0001		
3	0.3796	0.3828		0.0032		0.0032	3,093.42	< .0001	42.75	< .0001		
4	0.4205	0.4212	0.4216	0.0007	0.0004	0.0011	3,272.28	< .0001	10.54	0.0012	5.22	0.022
5	0.3680	0.3716		0.0036		0.0036	3,524.12	< .0001	54.84	< .0001		
6	0.3940	0.3966		0.0026		0.0026	2,945.72	< .0001	33.74	< .0001		
7	0.3839	0.3910	0.3917	0.0071	0.0007	0.0078	3,018.84	< .0001	92.40	< .0001	9.58	0.002
8	0.3405	0.3463		0.0058		0.0058	1,549.42	< .0001	40.49	< .0001		
9	0.3715	0.3759		0.0044		0.0044	1,475.10	< .0001	28.64	< .0001		
10	0.3805	0.3841		0.0036		0.0036	2,670.46	< .0001	42.32	< .0001		
11	0.3848	0.3880		0.0032		0.0032	3,013.79	< .0001	41.59	< .0001		
12	0.3576	0.3593		0.0017		0.0017	1,899.02	< .0001	14.54	0.0001		
13	0.3715	0.3784		0.0069		0.0069	2,225.63	< .0001	67.29	< .0001		
14	0.3969	0.4000		0.0031		0.0031	2,426.20	< .0001	32.43	< .0001		
15	0.3603	0.3678		0.0075		0.0075	3,650.96	< .0001	121.29	< .0001		
16	0.3873	0.3893		0.0020		0.0020	890.97	< .0001	7.46	0.0063		
17	0.3628	0.3660		0.0032		0.0032	716.08	< .0001	10.39	0.0013		
18	0.4164	0.4177		0.0013		0.0013	1,927.13	< .0001	10.33	0.0013		
19	0.3593	0.3669	0.3675	0.0076	0.0006	0.0082	2,110.03	< .0001	70.44	< .0001	5.38	0.020
20	0.3687	0.3759		0.0072		0.0072	844.48	< .0001	26.44	< .0001		
21	0.3951	0.3955		0.0004		0.0004	2,877.81	< .0001	5.17	0.0229		
22	0.4542	0.4592		0.0050		0.0050	2,850.71	< .0001	58.62	< .0001		
23	0.3576	0.3674		0.0098		0.0098	2,925.66	< .0001	125.82	< .0001		
24	0.3622	0.3628		0.0006		0.0006	2,715.21	< .0001	7.40	0.0065		

(Table continues)

Table D1 (continued)

Prompt	R^2 changes						$\chi2$ test for added terms					
	R^2 values			R^2 effect size			Ability (A)		Group (G)		A*G	
no.	Ability	Group	A*G	Uni	Non	Total	$\chi2$	*p*	$\chi2$	*p*	$\chi2$	*p*
25	0.3628	0.3651		0.0023		0.0023	2,615.96	< .0001	25.95	< .0001		
26	0.4324	0.4355		0.0031		0.0031	2,453.73	< .0001	31.72	< .0001		
27	0.3507	0.3536		0.0029		0.0029	2,809.85	< .0001	37.41	< .0001		
28	0.395	0.3982		0.0032		0.0032	2,091.68	< .0001	28.76	< .0001		
29	0.3937	0.3946		0.0009		0.0009	1,816.96	< .0001	6.82	0.009		
30	0.3692	0.3760		0.0068		0.0068	2,633.02	< .0001	78.11	< .0001		
31	0.3680	0.3737	0.3743	0.0057	0.0006	0.0063	2,477.88	< .0001	62.22	< .0001	5.96	0.015
32	0.3352	0.3451	0.3461	0.0099	0.0010	0.0109	2,083.30	< .0001	93.61	< .0001	9.62	0.002
33	0.3502	0.3609	0.3615	0.0107	0.0006	0.0113	3,443.61	< .0001	164.97	< .0001	8.12	0.004
34	0.3585	0.3679		0.0094		0.0094	3,063.50	< .0001	128.02	< .0001		
35	0.3938	0.3961		0.0023		0.0023	2,951.49	< .0001	28.84	< .0001		
36	0.3448	0.3516		0.0068		0.0068	1,950.38	< .0001	58.95	< .0001		
37	0.3609	0.3701		0.0092		0.0092	2,420.95	< .0001	98.78	< .0001		
38	0.3794	0.3830		0.0036		0.0036	3,528.59	< .0001	54.65	< .0001		
39	0.3518	0.3625		0.0107		0.0107	2,233.06	< .0001	106.37	< .0001		
40	0.3617	0.3709		0.0092		0.0092	2,567.67	< .0001	102.44	< .0001		
41	0.3677	0.3773		0.0096		0.0096	3,130.55	< .0001	133.64	< .0001		
42	0.4036	0.4068		0.0032		0.0032	3,200.38	< .0001	43.30	< .0001		
43	0.3693	0.3753	0.3761	0.0060	0.0008	0.0068	2,812.70	< .0001	74.15	< .0001	9.36	0.002
44	0.3497	0.3574	0.358	0.0077	0.0006	0.0083	3,215.71	< .0001	110.68	< .0001	7.25	0.007
45	0.3725	0.3781	0.3785	0.0056	0.0004	0.0060	2,760.46	< .0001	67.34	< .0001	4.89	0.027
46	0.3666	0.3688	0.3695	0.0022	0.0007	0.0029	3,361.33	< .0001	32.42	< .0001	9.02	0.003
47	0.3674	0.3747		0.0073		0.0073	4,230.69	< .0001	133.86	< .0001		
48	0.3691	0.3744		0.0053		0.0053	2,062.92	< .0001	46.87	< .0001		
49	0.3843	0.3878		0.0035		0.0035	4,123.45	< .0001	61.69	< .0001		
50	0.3681	0.3728		0.0047		0.0047	3,734.68	< .0001	77.18	< .0001		
51	0.3880						2,975.01	< .0001				
52	0.3800	0.3880		0.0080		0.0080	3,066.95	< .0001	105.59	< .0001		
53	0.3547	0.3590		0.0043		0.0043	2,287.13	< .0001	43.54	< .0001		
54	0.3853	0.3867	0.3876	0.0014	0.0009	0.0023	2,422.76	< .0001	14.96	0.0001	8.55	0.004
55	0.3584	0.3667		0.0083		0.0083	2,578.85	< .0001	94.56	< .0001		
56	0.3884	0.3990		0.0106		0.0106	2,666.46	< .0001	122.63	< .0001		
57	0.3793	0.3870		0.0077		0.0077	3,419.51	< .0001	113.17	< .0001		
58	0.3808	0.3847		0.0039		0.0039	3,214.49	< .0001	54.39	< .0001		
59	0.3719	0.3777		0.0058		0.0058	3,882.87	< .0001	97.62	< .0001		
60	0.3483	0.3527		0.0044		0.0044	2,128.36	< .0001	41.62	< .0001		
61	0.3308	0.3372	0.3377	0.0064	0.0005	0.0069	2,896.47	< .0001	86.03	< .0001	7.15	0.008
62	0.3879	0.3918		0.0039		0.0039	2,126.67	< .0001	35.77	< .0001		
63	0.3755	0.3793	0.3802	0.0038	0.0009	0.0047	2,918.29	< .0001	47.78	< .0001	11.45	0.001
64	0.3563	0.3617		0.0054		0.0054	2,370.29	< .0001	55.78	< .0001		
65	0.3553	0.3623		0.007		0.0070	3,050.68	< .0001	93.99	< .0001		

(Table continues)

Table D1 (continued)

Prompt no.	R^2 changes						$\chi 2$ Test for added terms					
	R^2 values			R^2 effect size			Ability (A)		Group (G)		A*G	
	Ability	Group	A*G	Uni	Non	Total	$\chi 2$	p	$\chi 2$	p	$\chi 2$	P
66	0.3459	0.3511		0.0052		0.0052	2,915.43	< .0001	68.59	< .0001		
67	0.3758	0.377		0.0012		0.0012	3,173.33	< .0001	16.15	< .0001		
68	0.3420	0.3505	0.3513	0.0085	0.0008	0.0093	2,394.46	< .0001	92.34	< .0001	8.82	0.003
69	0.3739	0.3771		0.0032		0.0032	2,960.36	< .0001	40.89	< .0001		
70	0.3841	0.3876		0.0035		0.0035	3,042.33	< .0001	45.77	< .0001		
71	0.3403	0.3433		0.003		0.0030	2,696.47	< .0001	35.89	< .0001		
72	0.3460	0.3475		0.0015		0.0015	3,709.32	< .0001	24.84	< .0001		
73	0.3715	0.3726		0.0011		0.0011	2,745.76	< .0001	12.21	0.0005		
74	0.3724	0.3749		0.0025		0.0025	2,674.89	< .0001	29.21	< .0001		
75	0.3829	0.3880		0.0051		0.0051	3,135.76	< .0001	68.02	< .0001		
76	0.3610	0.3640		0.003		0.0030	2,215.82	< .0001	28.77	< .0001		
77	0.3683	0.3741	0.3746	0.0058	0.0005	0.0063	2,876.25	< .0001	72.79	< .0001	6.47	0.011
78	0.3827	0.3868		0.0041		0.0041	2,382.98	< .0001	42.45	< .0001		
79	0.4170	0.4214	0.4220	0.0044	0.0006	0.0050	2,706.68	< .0001	49.69	< .0001	6.92	0.009
80	0.3598	0.3675		0.0077		0.0077	1,825.88	< .0001	61.89	< .0001		
81	0.3914	0.3927		0.0013		0.0013	2,603.78	< .0001	14.39	0.0001		
82	0.3640	0.3659		0.0019		0.0019	2,793.72	< .0001	22.67	< .0001		
83	0.3814	0.3845		0.0031		0.0031	1,878.70	< .0001	25.27	< .0001		
84	0.3715	0.3734		0.0019		0.0019	1,734.58	< .0001	14.17	0.0002		
85	0.3564	0.3640		0.0076		0.0076	1,966.85	< .0001	66.65	< .0001		
86	0.3909	0.3934		0.0025		0.0025	1,983.93	< .0001	21.20	< .0001		
87	0.3482	0.3507	0.3513	0.0025	0.0006	0.0031	3,446.19	< .0001	38.94	< .0001	7.98	0.005

Table D2

Intercept and Slope Parameters for the Logistic Regression for 87 Prompts (Phase II)

Prompt no.	Intercepts										Slopes		
	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	–6.18	–5.40	–3.93	–3.02	–1.54	–0.58	0.80	1.78	2.82	3.88	–0.51	–0.23	
2	–6.00	–4.91	–3.61	–2.56	–1.12	–0.06	1.30	2.23	3.21	4.43	–0.45	–0.52	
3	–5.80	–4.87	–3.68	–2.66	–1.39	–0.33	0.90	1.88	2.78	4.05	–0.52	–0.26	
4	–6.57	–5.63	–4.23	–3.12	–1.64	–0.57	0.76	1.81	2.74	3.91	–0.62	–0.14	0.04
5	–6.20	–5.29	–3.89	–2.84	–1.43	–0.34	1.01	2.03	2.95	4.14	–0.52	–0.27	
6	–6.15	–5.07	–3.74	–2.73	–1.34	–0.29	0.95	1.96	2.87	3.99	–0.54	–0.24	
7	–6.41	–5.30	–3.99	–2.94	–1.50	–0.39	1.00	2.02	3.01	4.17	–0.60	–0.38	0.05
8	–6.72	–5.37	–4.00	–3.09	–1.65	–0.58	0.88	1.90	2.96	4.06	–0.49	–0.34	
9	–6.38	–5.65	–3.96	–2.92	–1.57	–0.51	0.78	1.83	2.73	3.90	–0.52	–0.30	
10	–5.78	–4.76	–3.58	–2.49	–1.16	–0.05	1.17	2.13	3.09	4.28	–0.53	–0.28	

(Table continues)

Table D2 (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)
11	–5.88	–4.85	–3.72	–2.74	–1.39	–0.29	1.05	2.09	3.07	4.27	–0.53	–0.26	
12	–6.19	–5.46	–4.06	–3.10	–1.45	–0.44	0.94	1.95	2.99	4.11	–0.51	–0.18	
13	–6.26	–5.44	–4.09	–3.20	–1.70	–0.59	1.02	2.01	3.15	4.29	–0.51	–0.38	
14	–5.80	–4.92	–3.92	–2.88	–1.47	–0.38	1.04	2.03	3.03	4.12	–0.54	–0.26	
15	–5.67	–4.89	–3.55	–2.60	–1.15	–0.10	1.33	2.33	3.30	4.58	–0.50	–0.39	
16	–5.97	–5.32	–4.18	–3.08	–1.59	–0.59	0.76	1.73	2.76	4.02	–0.53	–0.20	
17	–7.30	–5.68	–4.38	–3.46	–1.87	–0.78	0.74	1.73	2.83	3.97	–0.53	–0.26	
18	–6.08	–5.36	–4.05	–2.98	–1.42	–0.41	0.93	1.84	2.89	3.92	–0.56	–0.17	
19	–5.61	–5.00	–3.91	–2.87	–1.41	–0.41	1.03	2.01	3.03	4.15	–0.56	–0.39	0.04
20	–5.46	–4.80	–3.74	–2.99	–1.47	–0.54	1.08	2.00	3.02	4.20	–0.52	–0.39	
21	–6.57	–5.43	–4.14	–3.09	–1.64	–0.58	0.70	1.69	2.67	3.86	–0.54	–0.09	
22	–5.69	–4.88	–3.72	–2.74	–1.28	–0.17	1.25	2.28	3.25	4.36	–0.60	–0.34	
23	–5.84	–4.73	–3.37	–2.40	–1.03	0.05	1.43	2.40	3.37	4.62	–0.50	–0.44	
24	–6.43	–5.41	–4.11	–3.12	–1.76	–0.71	0.65	1.61	2.57	3.71	–0.50	–0.11	
25	–6.33	–5.44	–4.27	–3.15	–1.62	–0.53	0.83	1.88	2.85	4.03	–0.51	–0.21	
26	–5.87	–5.07	–3.80	–2.84	–1.42	–0.46	0.96	1.98	2.99	4.16	–0.58	–0.27	
27	–6.36	–5.40	–4.14	–3.21	–1.67	–0.59	0.88	1.91	2.89	4.03	–0.50	–0.24	
28	–5.94	–4.88	–3.82	–2.74	–1.34	–0.23	1.00	2.05	2.97	4.16	–0.55	–0.26	
29	–5.67	–4.71	–3.55	–2.51	–1.22	–0.15	1.06	2.02	2.87	4.09	–0.53	–0.14	
30	–5.88	–4.88	–3.60	–2.60	–1.20	–0.17	1.09	2.12	3.13	4.22	–0.51	–0.37	
31	–5.74	–5.03	–3.94	–2.93	–1.42	–0.34	1.00	1.97	2.92	4.03	–0.58	–0.34	0.04
32	–5.81	–4.97	–3.86	–2.89	–1.45	–0.32	1.14	2.15	3.18	4.45	–0.55	–0.43	0.05
33	–5.90	–4.99	–3.70	–2.71	–1.35	–0.28	1.14	2.17	3.19	4.38	–0.55	–0.46	0.04
34	–6.02	–5.01	–3.63	–2.63	–1.27	–0.18	1.08	2.13	3.05	4.18	–0.51	–0.43	
35	–6.28	–5.16	–4.05	–2.95	–1.53	–0.46	0.84	1.89	2.86	4.12	–0.55	–0.22	
36	–6.03	–5.10	–3.69	–2.69	–1.29	–0.20	1.13	2.14	3.12	4.22	–0.49	–0.36	
37	–6.01	–4.95	–3.52	–2.59	–1.24	–0.16	1.20	2.26	3.19	4.43	–0.51	–0.43	
38	–6.02	–5.11	–3.92	–2.90	–1.57	–0.55	0.84	1.87	2.84	4.02	–0.52	–0.27	
39	–6.02	–4.96	–3.56	–2.53	–1.16	–0.08	1.20	2.28	3.28	4.47	–0.50	–0.46	
40	–6.07	–5.10	–3.69	–2.59	–1.15	–0.09	1.14	2.16	3.08	4.21	–0.51	–0.43	
41	–5.50	–4.64	–3.25	–2.23	–0.90	0.15	1.43	2.45	3.43	4.52	–0.52	–0.44	
42	–5.90	–4.94	–3.70	–2.77	–1.32	–0.24	1.02	2.06	2.98	4.14	–0.55	–0.26	
43	–5.75	–4.96	–3.70	–2.68	–1.29	–0.20	1.10	2.13	3.12	4.33	–0.58	–0.35	0.05
44	–5.38	–4.62	–3.37	–2.38	–1.05	–0.07	1.21	2.22	3.17	4.28	–0.54	–0.39	0.04
45	–5.70	–4.84	–3.62	–2.58	–1.19	–0.09	1.22	2.23	3.19	4.43	–0.57	–0.33	0.04
46	–5.79	–5.10	–3.90	–2.85	–1.50	–0.47	0.78	1.79	2.69	3.88	–0.57	–0.21	0.04
47	–5.59	–4.83	–3.60	–2.62	–1.23	–0.17	1.10	2.11	3.04	4.22	–0.51	–0.38	
48	–6.11	–4.97	–3.69	–2.69	–1.33	–0.27	1.10	2.09	3.10	4.27	–0.51	–0.33	
49	–6.21	–5.34	–3.95	–2.93	–1.49	–0.41	0.94	2.00	2.97	4.07	–0.53	–0.27	
50	–5.70	–5.00	–3.84	–2.90	–1.42	–0.36	1.02	2.03	2.99	4.12	–0.51	–0.31	
51	–6.19	–5.43	–4.14	–3.09	–1.71	–0.71	0.58	1.55	2.45	3.52	–0.53	0.00	
52	–5.81	–4.87	–3.81	–2.71	–1.25	–0.15	1.19	2.23	3.17	4.35	–0.53	–0.41	

(Table continues)

Table D2 (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)
53	–5.60	–4.91	–3.73	–2.79	–1.37	–0.33	0.98	1.98	2.99	4.09	–0.51	–0.29	
54	–6.27	–5.30	–3.95	–2.97	–1.54	–0.45	0.89	1.89	2.82	3.90	–0.63	–0.18	0.05
55	–5.91	–4.90	–3.57	–2.55	–1.05	–0.02	1.24	2.25	3.19	4.35	–0.51	–0.41	
56	–5.25	–4.54	–3.43	–2.44	–1.06	–0.04	1.32	2.33	3.35	4.51	–0.54	–0.47	
57	–5.78	–5.08	–3.69	–2.70	–1.25	–0.12	1.21	2.25	3.22	4.32	–0.52	–0.40	
58	–5.78	–4.95	–3.66	–2.56	–1.19	–0.15	1.11	2.12	3.03	4.21	–0.53	–0.28	
59	–5.69	–4.91	–3.72	–2.76	–1.33	–0.31	0.99	2.01	2.96	4.07	–0.52	–0.34	
60	–5.93	–5.22	–3.87	–2.83	–1.48	–0.44	0.91	1.96	3.05	4.20	–0.49	–0.29	
61	–5.95	–5.15	–3.77	–2.81	–1.41	–0.32	1.10	2.17	3.15	4.36	–0.53	–0.35	0.04
62	–5.93	–4.98	–3.81	–2.80	–1.51	–0.46	0.89	1.92	2.91	4.11	–0.53	–0.29	
63	–6.13	–5.39	–3.95	–2.96	–1.48	–0.40	1.03	2.04	3.04	4.20	–0.60	–0.28	0.05
64	–5.52	–4.69	–3.47	–2.46	–1.15	–0.07	1.05	2.05	2.93	4.07	–0.50	–0.33	
65	–6.01	–5.09	–3.77	–2.72	–1.32	–0.23	1.05	2.11	3.08	4.20	–0.50	–0.37	
66	–6.14	–5.34	–4.07	–3.09	–1.62	–0.50	0.90	1.90	2.98	4.17	–0.49	–0.32	
67	–5.95	–5.03	–3.80	–2.72	–1.41	–0.33	0.92	1.87	2.77	3.94	–0.52	–0.15	
68	–5.81	–5.12	–3.79	–2.74	–1.26	–0.20	1.25	2.25	3.28	4.37	–0.56	–0.41	0.05
69	–6.30	–5.31	–3.88	–2.87	–1.50	–0.45	0.97	1.95	2.87	4.01	–0.51	–0.25	
70	–5.98	–5.19	–3.83	–2.84	–1.38	–0.31	1.03	2.03	2.95	4.18	–0.54	–0.27	
71	–5.79	–4.99	–3.90	–2.91	–1.50	–0.38	1.03	2.05	3.04	4.22	–0.48	–0.24	
72	–6.17	–5.29	–3.99	–2.93	–1.47	–0.40	0.93	1.96	2.91	4.05	–0.48	–0.17	
73	–6.32	–5.34	–3.84	–2.78	–1.46	–0.41	0.79	1.71	2.57	3.61	–0.51	–0.14	
74	–5.78	–5.10	–3.79	–2.79	–1.40	–0.32	1.03	2.02	2.96	4.02	–0.52	–0.23	
75	–5.74	–4.86	–3.64	–2.64	–1.20	–0.11	1.16	2.22	3.11	4.21	–0.53	–0.33	
76	–6.84	–5.94	–4.38	–3.35	–1.75	–0.67	0.87	1.87	2.92	3.96	–0.50	–0.24	
77	–6.19	–5.34	–3.96	–2.91	–1.50	–0.41	0.89	1.90	2.88	4.01	–0.57	–0.35	0.04
78	–5.61	–4.84	–3.50	–2.51	–1.15	–0.03	1.32	2.34	3.32	4.54	–0.53	–0.29	
79	–6.00	–4.80	–3.53	–2.49	–1.16	–0.03	1.24	2.31	3.25	4.43	–0.63	–0.31	0.04
80	–5.61	–4.84	–3.60	–2.71	–1.20	–0.12	1.30	2.30	3.38	4.56	–0.51	–0.39	
81	–5.92	–4.92	–3.60	–2.65	–1.34	–0.25	0.98	2.00	2.95	4.14	–0.55	–0.16	
82	–6.18	–5.49	–4.26	–3.19	–1.76	–0.64	0.77	1.77	2.75	3.89	–0.50	–0.19	
83	–6.09	–5.25	–3.82	–2.83	–1.34	–0.30	1.10	2.17	3.10	4.24	–0.53	–0.25	
84	–6.14	–5.39	–4.16	–3.12	–1.63	–0.58	0.94	1.96	3.09	4.32	–0.52	–0.20	
85	–6.07	–5.11	–3.67	–2.86	–1.24	–0.26	1.23	2.20	3.22	4.38	–0.50	–0.39	
86	–6.10	–5.27	–4.00	–2.94	–1.44	–0.42	0.96	1.98	3.01	4.16	–0.53	–0.23	
87	–5.86	–5.09	–3.78	–2.77	–1.42	–0.33	1.03	2.05	3.03	4.14	–0.55	–0.22	0.04

Appendix E

Prompt Difficulty Indices

Table E1 shows the prompt difficulty indices estimated based on the logistic regression Step 1 model:

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

Table E1

Mean Raw Essay Scores and Prompt Difficulty Indices for 47 Prompts in Phase I and 87 Prompts in Phase II

Prompt no.	Phase I				Phase II			
	N	Mean	$\bar{\xi}$	$1/\bar{\xi}$	N	Mean	$\bar{\xi}$	$1/\bar{\xi}$
1	*	*	*	*	5,754	4.11	2.91	0.34
2	*	*	*	*	2,591	4.10	3.23	0.31
3	2,961	4.05	2.37	0.42	8,333	4.04	2.47	0.40
4	2,970	4.10	2.54	0.39	8,019	4.09	2.54	0.39
5	2,936	4.06	2.80	0.36	9,868	4.06	2.66	0.38
6	2,805	4.03	2.49	0.40	7,641	4.04	2.43	0.41
7	*	*	*	*	8,066	4.09	2.99	0.33
8	*	*	*	*	4,647	4.17	3.39	0.29
9	2,576	4.20	3.17	0.32	4,059	4.16	3.11	0.32
10	2,507	3.99	2.14	0.47	7,217	3.96	2.11	0.47
11	*	*	*	*	8,078	3.99	2.30	0.43
12	*	*	*	*	5,470	4.03	2.66	0.38
13	*	*	*	*	6,164	4.14	3.21	0.31
14	*	*	*	*	6,304	4.03	2.37	0.42
15	2,655	3.97	2.38	0.42	10,312	3.99	2.41	0.42
16	*	*	*	*	2,405	4.08	2.70	0.37
17	*	*	*	*	2,066	4.19	3.43	0.29
18	*	*	*	*	4,739	4.04	2.34	0.43
19	*	*	*	*	6,005	4.13	2.94	0.34
20	*	*	*	*	2,325	4.11	2.77	0.36

(Table continues)

Table E1 (continued)

Prompt no.	Phase I				Phase II			
	N	Mean	$\overline{\xi}$	$1/\overline{\xi}$	*N*	Mean	$\overline{\xi}$	$1/\overline{\xi}$
21	2,971	4.05	2.61	0.38	7,499	4.06	2.58	0.39
22	*	*	*	*	6,453	4.00	2.05	0.49
23	2,926	3.98	2.24	0.45	8,296	3.99	2.39	0.42
24	*	*	*	*	7,692	4.08	2.90	0.34
25	*	*	*	*	7,407	4.07	2.92	0.34
26	*	*	*	*	5,801	4.04	2.29	0.44
27	2,400	4.13	2.98	0.34	8,326	4.11	3.05	0.33
28	*	*	*	*	5,424	4.03	2.27	0.44
29	2,438	3.94	1.91	0.52	4,684	3.92	1.82	0.55
30	2,893	3.99	2.33	0.43	7,298	4.06	2.57	0.39
31	2,293	4.14	3.01	0.33	6,951	4.11	2.76	0.36
32	*	*	*	*	6,369	4.07	3.05	0.33
33	2,753	4.13	3.05	0.33	10,046	4.10	2.96	0.34
34	2,634	4.09	2.81	0.36	8,779	4.11	2.88	0.35
35	2,318	4.07	2.38	0.42	7,703	4.07	2.55	0.39
36	*	*	*	*	5,785	4.04	2.80	0.36
37	2,393	4.06	2.75	0.36	6,885	4.06	2.69	0.37
38	2,575	4.17	2.72	0.37	9,521	4.12	2.76	0.36
39	2,047	4.11	2.77	0.36	6,482	4.08	2.78	0.36
40	2,444	4.08	2.77	0.36	7,180	4.10	2.82	0.35
41	2,727	3.95	2.12	0.47	8,825	3.95	2.12	0.47
42	2,035	4.07	2.32	0.43	8,131	4.04	2.26	0.44
43	*	*	*	*	7,846	4.03	2.51	0.40
44	*	*	*	*	9,391	3.97	2.39	0.42
45	*	*	*	*	7,629	3.95	2.27	0.44
46	4,286	4.08	2.72	0.37	9,390	4.08	2.67	0.37
47	4,071	4.10	2.74	0.36	11,760	4.06	2.60	0.39
48	*	*	*	*	5,691	4.02	2.61	0.38
49	3,787	4.11	2.86	0.35	10,975	4.08	2.69	0.37
50	3,229	4.10	2.83	0.35	10,382	4.07	2.69	0.37
51	3,685	4.07	2.61	0.38	7,771	4.07	2.49	0.40
52	4,187	4.04	2.58	0.39	8,279	4.07	2.57	0.39
53	*	*	*	*	6,641	4.06	2.56	0.39
54	*	*	*	*	6,476	4.06	2.45	0.41
55	*	*	*	*	7,374	4.01	2.54	0.39

(Table continues)

Table E1 (continued)

Prompt no.	Phase I				Phase II			
	N	Mean	$\bar{\xi}$	$1/\bar{\xi}$	*N*	Mean	$\bar{\xi}$	$1/\bar{\xi}$
56	*	*	*	*	7,107	4.02	2.26	0.44
57	3,481	4.06	2.81	0.36	9,193	4.04	2.59	0.39
58	3,756	3.99	2.43	0.41	8,675	3.99	2.28	0.44
59	3,894	4.08	2.66	0.38	10,694	4.08	2.66	0.38
60	*	*	*	*	6,258	4.05	2.82	0.35
61	3,391	4.06	2.83	0.35	9,043	4.06	2.91	0.34
62	*	*	*	*	5,603	4.07	2.62	0.38
63	*	*	*	*	8,018	4.04	2.68	0.37
64	3,976	4.06	2.45	0.41	6,698	4.03	2.42	0.41
65	3,188	4.04	2.73	0.37	8,699	4.08	2.85	0.35
66	2,747	4.18	3.34	0.30	8,633	4.13	3.16	0.32
67	3,813	3.96	2.24	0.45	8,646	4.00	2.30	0.43
68	3,078	4.07	2.74	0.37	7,208	4.06	2.80	0.36
69	3,109	4.09	2.92	0.34	8,124	4.07	2.78	0.36
70	*	*	*	*	8,168	4.05	2.48	0.40
71	*	*	*	*	8,113	4.00	2.63	0.38
72	2,598	4.06	2.61	0.38	11,011	4.02	2.66	0.38
73	3,285	4.04	2.64	0.38	7,426	4.05	2.66	0.38
74	2,836	3.98	2.41	0.42	7,330	4.01	2.37	0.42
75	3,191	3.97	2.35	0.43	8,316	3.98	2.31	0.43
76	*	*	*	*	6,321	4.13	3.36	0.30
77	3,083	4.11	2.84	0.35	7,943	4.15	3.07	0.33
78	*	*	*	*	6,369	3.89	1.97	0.51
79	2,654	3.95	1.90	0.53	6,674	3.97	1.99	0.50
80	*	*	*	*	5,233	4.00	2.42	0.41
81	2,764	3.96	2.08	0.48	6,843	3.99	2.01	0.50
82	2,684	4.12	3.11	0.32	7,785	4.11	3.02	0.33
83	*	*	*	*	5,111	3.99	2.42	0.41
84	*	*	*	*	4,767	4.05	2.63	0.38
85	*	*	*	*	5,649	4.04	2.80	0.36
86	*	*	*	*	5,186	4.03	2.55	0.39
87	2,580	4.03	2.56	0.39	10,197	3.98	2.46	0.41

Note. Asterisks (*) in some of the cells mean that no examinee data for the particular prompt were available in the Phase I study.

Appendix F

Taxonomy of TOEFL CBT Writing Prompts (Phase I)

Table F1

TOEFL CBT Writing Prompts (Phase I)

Prompt no.	Difficulty- $1/\bar{\xi}$	Gender diff. (RBES)	Topic	Word count	Low freq. words	Personal exper. probab.	TOEFL class.
66	0.30	0.11	Housing	24	0	Medium	TP/ERD
9	0.32	0.13	School	27	0	High	I/ERD
82	0.32	0.05	School	35	0	Medium	TP/ERE
33	0.33	0.18	Friends	47	0	High	CC/TP/ER
31	0.33	0.16	Customs	40	0	Low	CC/TP/ED
27	0.34	0.09	Travel	38	0	High	CC/TP/ER
69	0.34	0.19	School	32	0	High	A/I/ERD
49	0.35	0.13	Movies	32	0	High	AD/ERE
77	0.35	0.18	School	39	0	Medium	TP/ERD
61	0.35	0.21	Housing	34	0	High	I/ERE
50	0.35	0.14	School	44	0	High	TP/ERD
34	0.36	0.22	Children	48	0	High	CC/TP
57	0.36	0.19	School	42	0	High	TP/ERE
	0.36	0.10	Hometown	23	1	High	I/ERE
39	0.36	0.25	Friends	49	0	High	CC/TP
40	0.36	0.17	Change	40	0	Medium	CC/TP
37	0.36	0.21	Planning	61	0	High	CC/TP/ERE
47	0.36	0.21	Art/Music	29	0	High	AD/ER
68	0.37	0.21	Leisure	44	2	High	TP/ERE
65	0.37	0.20	Vacations	40	0	High	CC/TP/ERE
46	0.37	0.13	Science	39	0	High	AD/ERE
38	0.37	0.14	School	44	0	High	I/D/EE
59	0.38	0.15	School	32	0	High	AD/ERE
73	0.38	0.11	Resource	40	0	Low	I/ERE
21	0.38	0.02	Space	58	1	Low	CC/TP/ERD
72	0.38	0.03	Animals	23	0	High	I/ERD
51	0.38	0.02	Research	49	0	Low	TP/ER
52	0.39	0.17	Children	28	0	High	D/+–/ERE
87	0.39	0.09	Housing	25	0	Low	TP/ER

(Table continues)

Table F1 (continued)

Prompt no.	Difficulty- $1/\bar{\xi}$	Gender diff. (RBES)	Topic	Word count	Low freq. words	Personal exper. probab.	TOEFL class.
4	0.39	0.07	Factory	37	0	Low	TP/+–
6	0.40	0.10	Media	17	0	Medium	A/D/ERE
64	0.41	0.13	Music	30	0	Low	A/D/ERE
58	0.41	0.13	Games	27	0	High	AD/ERE
74	0.42	0.13	Zoo	24	0	Medium	AD/ERD
15	0.42	0.18	Neighbors	27	0	High	I/EED
35	0.42	0.13	University	46	1	Low	CC/TP/+–
3	0.42	0.12	Books	34	0	Low	CC/TP/E
75	0.43	0.12	Smoking	38	0	Low	D/TP/ERE
30	0.43	0.15	Life	61	0	Low	CC/TP/ER
42	0.43	0.15	Decisions	46	0	Low	AD/ERE
23	0.45	0.22	Roommate	48	3	Medium	TP/ERD
67	0.45	0.05	Advertising	44	0	Low	TP/ERD
10	0.47	0.16	Success	34	0	Low	AD/ERE
41	0.47	0.19	Clothing	33	0	Low	AD/EE
81	0.48	0.08	Athletes	29	1	Low	D/TP/ERE
29	0.52	0.09	Barter	38	0	Low	CC/TP
79	0.53	0.17	Complain	38	0	Medium	D/TP/ERE

Note: Abbreviations in the table have the following meanings:

+/–	Advantages/Disadvantages
A	Analyze
AD	Agree/Disagree
CC	Compare/contrast
D	Discuss
E	Explain
ER	Explain using reasons
EE	Explain using examples
ED	Explain using details
ERD	Explain using reasons and details
ERE	Explain using reasons and examples
EED/EDE	Explain using details and examples
I	Identify
TP	Take a position

Appendix G

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).

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 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: 609-771-7100
Email: toefl@ets.org
Web site: www.ets.org/toefl

73313-035280 • Y24E.05 • Printed in U.S.A.
I.N. 990111