ISSN 1930-9317

TOEFL

TOEFL iBT Research Report

TOEFLiBT-09
July 2009

Does Content Knowledge Affect TOEFL iBT™ Reading Performance? A Confirmatory Approach to Differential Item Functioning

Ou Lydia Liu

Mary Schedl

Jeanne Malloy

Nan Kong

Listening.
Learning.
Leading.®

Does Content Knowledge Affect TOEFL iBT™ Reading Performance? A Confirmatory Approach to Differential Item Functioning

Ou Lydia Liu, Mary Schedl, Jeanne Malloy, and Nan Kong
ETS, Princeton, New Jersey

RR-09-29

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

Abstract

The TOEFL iBT™ has increased the length of the reading passages in the reading section compared to the passages on the TOEFL® computer-based test (CBT) to better approximate academic reading in North American universities, resulting in a reduced number of passages in the reading test. A concern arising from this change is whether the decrease in topic variety increases the likelihood that an examinee's familiarity with the particular content of a given passage will influence the examinee's reading performance. This study investigated differential item functioning and differential bundle functioning for six TOEFL iBT reading passages, three involving physical science and three involving cultural topics. The majority of items displayed little or no differential item functioning (DIF). When all of the items in a passage were examined, none of the passages showed differential functioning at the passage level. Hypotheses are provided for the DIF occurrences. Implications for fairness issues in test development are also discussed.

Key words: Content schemata, differential item functioning, differential bundle functioning, reading comprehension, TOEFL iBT

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. The test is now wholly owned and operated by ETS.

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, two-year colleges, and nonprofit educational exchange agencies.

❖ ❖ ❖

Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, TOEFL iBT. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research and technical reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

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

Current (2008-2009) members of the TOEFL Committee of Examiners are:

Alister Cumming (Chair)	University of Toronto
Geoffrey Brindley	Macquarie University
Frances A. Butler	Language Testing Consultant
Carol A. Chapelle	Iowa State University
John Hedgcock	Monterey Institute of International Studies
Barbara Hoekje	Drexel University
John M. Norris	University of Hawaii at Manoa
Pauline Rea-Dickins	University of Bristol
Steve Ross	Kwansei Gakuin University
Mikyuki Sasaki	Nagoya Gakuin University
Robert Schoonen	University of Amsterdam
Steven Shaw	University of Buffalo

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

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Acknowledgments

The research was funded by the TOEFL program at ETS. The opinions or interpretations expressed in the manuscript are those of the authors and not necessarily of the TOEFL program. The authors are thankful to Namrata Tognatta and Burton Fried for their assistance in survey design and data collection and to Neil Dorans for his suggestions on the standardization DIF method. The authors are also grateful to Jinghua Liu, Dan Eignor, Carsten Roever, Mikyung Kim Wolf, and John Young for their comments and edits on an earlier draft of this paper. Kim Fryer and Kathy Howell helped format and edit this report.

Table of Contents

List of Tables

Page

Executive Summary

The Test of English as a Foreign Language™ (TOEFL®) has been changed substantially in both its content and delivery format compared to the old TOEFL computer-based test (CBT), which has been phased out. In its current form as an Internet-based test (iBT), one of its most prominent changes is that the length of reading passages in the TOEFL iBT™ exam has increased from 350 words or fewer to approximately 700 words. This change marks a decrease in the number of passages contained in each form, leading to a decrease in topic variety. Concerns arise that the decreased topic variety may increase the likelihood that test takers' familiarity with the particular content of a given passage will influence their reading performance on the test.

This study examined whether outside knowledge, gained either through an examinee's academic major or from immersion in a particular culture, influences TOEFL iBT reading performance, and if so, to what extent. Six passages from five TOEFL iBT administrations between October 2007 and February 2008 were examined. Three of the passages focus on topics in physical science, and the rest emphasize European or Japanese cultures. A survey was sent to 58,038 test takers who took the TOEFL iBT during the five administrations. The survey included questions about test takers' major fields of study and cultural backgrounds. Responses were received from 8,692 test takers in total, and the number of respondents for each of the six passages ranged from 1,088 to 2,102.

Differential item functioning (DIF) and differential bundle functioning (DBF) were used to investigate the impact of outside knowledge on TOEFL iBT reading performance. DIF occurs for an item when differences in performance exist after examinees are matched on the abilities that the item is intended to measure. In this study, DIF analyses were used to examine performance differences between examinees who were hypothetically favored by their familiarity with the physical science topics or cultural topics and those who were not. The DIF analyses are supported by the rationale that the TOEFL iBT is a test of communicative language skills rather than of specific content knowledge, and therefore the test results should not be affected by test takers' major field of study or cultural background. In addition, items in a passage with certain common characteristics (e.g., presence of technical terminology) were bundled together to examine whether these items, as a whole, display differential bundle functioning (DBF). The effect size of the DIF and DBF was determined following the ETS DIF

guidelines and categorizations, with the A category suggesting no or little DIF, the B category suggesting small to moderate DIF, and the C category suggesting large DIF.

The majority of the items displayed little DIF. Very few items showed B- or C-DIF. The direction of the DIF occurrences was not consistent, with some items favoring the focal group and others favoring the reference group. Of the six item bundles examined, two showed differential functioning in favor of the focal group. Examination of the complete item set in a passage revealed no differential functioning at the passage level for any of the six passages. The DIF items and the items in the bundles that showed DBF will be closely examined to determine if the performance differences represent any bias related to group membership. Distinctions need to be drawn between DIF occurrence and item bias. The presence of DIF is *not* sufficient for item bias if no substantive factors exist that can explain the performance differences with regard to group membership. Items with large DIF values need to be carefully scrutinized to determine if they measure any ability irrelevant to the ability of interest. Items with large DIF values may not be selected for further use unless the items meet all required test specifications and fairness considerations. In summary, there is no consistent evidence that test takers with physical science knowledge or particular cultural background are favored by the content of the TOEFL iBT reading passages. For next steps of research, it would be worthwhile to investigate the interaction between test takers' language proficiency level and DIF occurrences, as the existence or magnitude of DIF may vary across test takers at different language proficiency levels.

The Test of English as a Foreign Language™ (TOEFL®) has undergone substantial changes in both its delivery platform and content. TOEFL has been transformed from a traditional paper and pencil test to a computer-based test (CBT) and now to an Internet based test (iBT). The design of the TOEFL iBT™ reading test, as described by the *TOEFL 2000 Reading Framework* (Enright et al., 2000), was guided by a reader-purpose perspective. Specifically, the reading items have been designed to reflect two academic reading purposes (Enright et al.): *reading for basic comprehension* and *reading to learn*. In turn, the length of each reading passage increased from 350 words or fewer in the paper and pencil and CBT tests to approximately 700 words in the iBT test. The change was supported by the rationale that longer passages can better approximate the academic reading load at North American universities. Longer passages also allow for the design of *reading to learn* items that are based on more substantive text content. Since the reading testing time remains unchanged, fewer passages are contained within the newer version of the test. A growing concern with these changes is that the decrease in topic variety may increase the likelihood that a test taker's familiarity with the particular content of a given passage will influence his or her reading performance.

Therefore, the purpose of this study was to determine whether prior knowledge, gained either through studies in an examinee's academic major or from his or her immersion in a particular culture, influences TOEFL iBT reading performance, and if so, to what extent. Differential item functioning (DIF) and differential bundle functioning (DBF) techniques were used to address this research question. Results from this study provide important evidence on construct validity of the TOEFL iBT.

Literature Review

Impact of Content Knowledge and Cultural Background on Reading Comprehension

Prior knowledge that may advantage certain test takers on reading passages comes from two primary sources: knowledge gained from systematic training in a major field of study and knowledge accumulated from being immersed in a specific culture. Content schema theory (e.g., Rumelhart, 1980) specifies that test takers who have acquired knowledge in a particular field develop schemata regarding that area, and their accumulated prior knowledge can facilitate the understanding of passages related to the field of their study. Note that the terminology, *prior knowledge*, *outside knowledge,* and *content schema* are used interchangeably in the following text.

In contrast to the abundant research on reading proficiency, there is relatively little empirical research on the effect of content schemata. The earliest study on the impact of content schemata on reading performance was conducted by Bartlett in 1932. He found that when English participants were asked to read a passage about an unfamiliar culture and to repeat the information in the passage, distortions occurred, reflecting the readers' past experience instead of what was present in the passage. Brown (1982) administered an engineering reading test consisting of three reading passages to 116 college students at UCLA. Results showed that engineering students performed better than nonengineering students on items involving both specific engineering knowledge and general engineering content. Erickson and Molloy (1983) conducted a similar study based on a reading test that was also administered to a group of 83 college students. They were able to confirm Brown's finding that engineers significantly outperformed nonengineers with regard to engineering content, in both specific and general engineering reading. Similar findings were also reported by Alderson and Urquhart (1984) that engineering students (n = 11) performed better on engineering-related reading passages and economics students (n = 11) performed better on economics-related passages.

The sample sizes of the above studies are relatively small or modest. Hale (1988) examined the impact of major-field area on reading performance with a larger sample. He examined data from 32,467 graduate school applicants from four TOEFL paper-and-pencil administrations and found that students in two key major-field groups, the humanities/social sciences and the biological/physical sciences, performed significantly better on passages involving content relevant to their majors than on other passages. Although differences were statistically significant, the practical effect sizes were found to be small. The sources of the text could account for the small effect sizes. Hale hypothesized that because the texts were drawn from general readings, the advantage of studying a particular major was not as great as it could have been had the texts been drawn from specialized textbooks.

Cultural influences can also affect test takers' reading performance. Keshavarz, Atai, and Ahmadi (2007) investigated the contribution of content and background knowledge, vocabulary and syntactic knowledge, and L2 proficiency to reading comprehension and recall. The participants were 240 male Iranian students who learned English as a foreign language. Each participant was tested with two types of texts: an extract from the biography of an Islamic religious leader who is supposed to be familiar to the Muslim participants, and an extract from

the biography of a non-Islamic religious figure. The authors found that familiarity with content was significantly correlated with reading comprehension test scores and recall scores ($p < .000$).

Floyd and Carrell (1987) designed an experimental study for 34 intermediate-level ESL students attending a college-level English program. Participants in the treatment group received two training sessions on cultural background knowledge. Pre- and post- culture-related reading tests were used to measure any potential change in reading ability for the treatment and control groups. The authors reported that students in the treatment group performed significantly better than those in the control group on passages containing pertinent cultural information. Chihara, Sakurai, and Oller (1989) also found that after culturally unfamiliar terms were altered to familiar ones on two reading passages, student performance was significantly improved among 159 Japanese junior college students. Similar findings were reported in a later study (Sasaki, 2000) that after unfamiliar words in a cloze test were changed to more familiar ones, students ($n = 30$) in the familiar group significantly outperformed the students ($n = 30$) in the unfamiliar group. Abu-Rabia (1996) conducted a study examining the effect of cultural background on student comprehension of familiar and unfamiliar information among 83 Israeli high school students. The students were tested with passages including three Jewish and three non-Jewish stories. The author found that students understood the culturally familiar stories significantly better than the unfamiliar ones.

Objective of This Study

The above synthesis provides evidence that familiarity with the content may have a positive impact on student reading performance. In the context of the TOEFL iBT, the present study aims to address one specific research question: Do content schemata have an impact on TOEFL iBT reading performance, after controlling for test takers' reading proficiency? Most previous studies have addressed the first part of this research question by using *t* tests or ANOVA to examine the performance differences between two groups or among multiple groups. However, results revealed in this way may relate to genuine differences in reading abilities between the compared groups rather than to content schemata. The phenomenon in which some test takers perform better on an item than others is referred to as *item impact*. Differential item functioning (DIF), on the other hand, suggests that performance differences exist after examinees are matched on the abilities measured by the test items (Holland & Wainer, 1993). Therefore, in the case of DIF, the score differences are unexpected performance differences between two

groups of test takers who are supposed to be comparable on the ability that the item is intended to measure (Dorans & Schmitt, 1993, p.138). In this study, DIF analyses were used to examine performance differences between examinees who are hypothetically favored by content schemata and those who are not. The DIF analyses are supported by the rationale that TOEFL iBT is a test of communicative language skills rather than of specific content knowledge, and thus the test results should not be affected by test takers' major field of study or cultural background.

By addressing the above question, this study aims to provide research evidence on the construct validity of the TOEFL iBT reading test. To our knowledge, no studies have been conducted to examine the impact of content schemata on the TOEFL reading test within the last 20 years, since the Hale (1988) study. The TOEFL reading test has undergone major changes since then, featuring longer reading passages and a greater variety of cultural topics. An up-to-date investigation is needed to provide timely information for test takers, test developers, and users of the TOEFL iBT scores (i.e., admission officers in universities). Results from this study have important implications for item development for the TOEFL iBT. Items identified with DIF shall be further scrutinized to determine if they should be retained, revised, or replaced to ensure a fair evaluation of all test takers. If the DIF occurrence is likely to be introduced by true differences in abilities of interest, then the existence of DIF may not violate the validity of the test. However, if the presence of DIF is related to some construct-irrelevant factors, the interpretation of the test scores may be inappropriate, thus threatening the validity of the interpretations (Kane, 2006). In a high-stakes testing situation such as TOEFL iBT, it is crucial to eliminate unwarranted item advantages for certain test takers, in order to ensure the validity of the test scores.

Method

Selection of TOEFL iBT Reading Passages

The authors worked closely with two experienced TOEFL iBT reading assessment developers in selecting passages for the DIF investigations. The selection was guided by three principles: (a) the passage should contain either heavily physical science-oriented content or be mainly about cultural knowledge from outside of the United States (e.g., Japanese arts); (b) the passage should come from one of the most recent TOEFL iBT administrations, no earlier than October 2007 (recency was considered to increase the likelihood of a test taker's response to a survey sent to them); and (c) the passage should have been taken by a fairly large number of test

takers (i.e., more than 10,000). The first criterion was imposed in the selection process because some of the items may display DIF, given their heavy emphasis on physical science or cultural content. In this sense, the selected passages do not represent typical TOEFL iBT reading passages, as other passages contain more neutral content than the selected ones.

Six reading passages were selected from the TOEFL iBT reading tests for examination of content schemata. These six passages were tested in five administrations between October 2007, and February 2008, including two passages tested in the same administration. Note that each test administration has three reading passages in total. Three of the selected passages for this study feature physical science topics, including one about planet formation, one about the planet Venus, and one about planetary impact events. Two of the six passages contain materials about Japan, one on landscape painting in medieval Japan, and the other on factors influencing the 13th-century restructuring of Japan. The sixth passage is about European academic art and the salon. It was hypothesized that the first three passages may favor test takers with physical science backgrounds, the two Japan-related passages may favor examinees familiar with East Asian culture, since these two passages also contain information about other East Asian countries, and the passage on European art and salon may advantage examinees familiar with European arts or history.

Each reading passage has 13 or 14 multiple choice items based on the content of the passage. All but one item in each set were dichotomously rated with a 0 or 1 score. The remaining item in each set was rated on a 0, 1, or 2 scale. For future reference in this paper, the passage identified for DIF analysis is referred to as the key passage, and the two other passages in the same administration are referred to as the nonkey passages. In total, there are six key passages. The number of items in each passage is provided in Table 1. The reliability indicated by Cronbach's alpha for the key passages for the reading tests is also provided in Table 1. The first column in this table includes the content of each key passage, followed by the number of items in each key passage, each bundle within the key passage, and the number of items in the entire reading test, which includes three passages. Item bundles are constructed for each passage on the basis of certain common item characteristics. For example, the item bundle for each physical science passage consists of items with heavy technical terminology. Content experts (i.e., item developers) hypothesized that items with heavy presence of technical terminology may favor test takers with a physical science background. The construction of item bundles is

discussed in detail in the following section. The last two columns are the Cronbach's alpha for the key passages and for the entire reading test, consisting of three passages. All the reliabilities are above .70, suggesting good internal consistency.

Table 1

Number of Items and Reliabilities

Passage	Number of items			Key passage reliability	Reading test reliability
	Key passage	Bundle	Reading test		
European art and salon	14	7	40	.77	.87
Restructuring of Japan	14	4	41	.71	.87
Landscape painting in Japan	14	8	41	.70	.87
Planet formation	14	6	40	.77	.88
Planet Venus	14	6	41	.78	.89
Planetary impact events	14	8	41	.72	.87

Note. The reading test is the test that contains the key passage under investigation.

Survey and Data Collection

A survey was designed to gather information about the test taker's major field of study and cultural background. The survey included demographic variables (i.e., gender, ethnicity, native language), academic variables (e.g., academic status when taking the TOEFL iBT, major field of study), cultural variables (e.g., the culture the examinee most identifies with), and test-related questions (e.g., the number of times taking the TOEFL iBT). For the DIF investigations in this study, the questions *If you were a college or graduate student when you took the TOEFL iBT, what was your major field of study?* and *Which culture do you most identify with?* are of key interest. Information about examinees' major field of study was used to investigate the impact of disciplinary schemata, and information about their cultural background was used to investigate the impact of cultural schemata.

The survey was sent to 58,038 test takers via e-mail (see Appendixes A and B for the e-mail message and the survey). After test takers clicked the link provided in the e-mail, they were directed to a Web page where they could complete the survey online. The survey was designed

so that respondents had to answer each question to proceed, so there were no missing data. These test takers were selected because (a) they took a TOEFL iBT containing reading passages of interest, (b) they provided an e-mail address during the test-taking, and (c) during the test-taking they authorized ETS to contact them for research purposes. Monetary incentives were provided to the first 300 respondents. Two weeks after the first contact, a follow-up e-mail was sent again to non-responsive examinees.

Sample

The response rate was 15%, and data were collected from 8,692 examinees. To carry out the DIF analysis, respondents were divided into a focal and a reference group for each passage. The focal group is the group under study, and the reference group is often used as a reference point for comparison. These terms originated with Holland and Thayer (1988) and Holland (1985). The determination of focal and reference group membership in this study varied across passages. For the three passages involving physical science content, examinees identifying themselves as physical science majors were considered focal group members, and the rest were categorized in the reference group. For the two passages involving Japanese arts, respondents identifying themselves as most familiar with East Asian culture were treated as focal group members, and the rest of the respondents were classified in the reference group. East Asian culture was deemed appropriate for these two passages, because besides information on Japanese arts, they also contained general information about other countries in East Asia (e.g., China). For the passage involving European academic art and the salon, respondents who identified themselves as most familiar with Eastern European, Western European, or Scandinavian culture were considered focal group members, and the rest of the examinees were categorized into the reference group. The demographic summary of the respondents in both focal and reference groups is provided in Table 2. Information on other variables such as gender and language group was gathered to ensure a balanced sample for this study. These variables were not used to conduct subgroup analysis due to insufficient sample sizes in the focal group.

Differential Item Functioning

Several methods have been widely applied for DIF research in language testing during the past 15 years (Ferne & Rupp, 2007), including the standardization approach (Dorans & Holland, 1993; Dorans & Kulick, 1983). Besides its application in educational research, the standardization

Table 2

Gender and Academic Status of Test Takers in the Focal and Reference Groups

Key passages	Male (%)	Female (%)	High school students (%)	College students (%)	Graduate students (%)	Others (%)	Total (*N*)
				Focal group			
European art and salon	207 (46)	246 (54)	52 (11)	198 (44)	145 (32)	58 (13)	453
Restructuring of Japan	206 (48)	223 (52)	137 (32)	144 (34)	109 (25)	39 (09)	429
Landscape painting in Japan	123 (39)	191 (61)	81 (26)	109 (35)	91 (29)	33 (11)	314
Planet formation	195 (72)	74 (28)	0 [a]	139 (52)	130 (48)	0	269
Planet Venus	139 (70)	60 (30)	0	105 (53)	94 (47)	0	199
Planetary impact events	257 (73)	94 (27)	0	163 (46)	188 (54)	0	351
				Reference group			
European art and salon	904 (56)	713 (44)	165 (10)	760 (47)	536 (33)	156 (10)	1,617
Restructuring of Japan	842 (50)	831 (50)	260 (16)	600 (36)	613 (37)	200 (12)	1,673
Landscape painting in Japan	812 (53)	730 (47)	210 (14)	544 (35)	605 (39)	183 (12)	1,542
Planet formation	546 (48)	600 (52)	216 (19)	369 (32)	376 (33)	185 (16)	1,146
Planet Venus	352 (40)	537 (60)	261 (29)	240 (27)	261 (29)	127 (14)	889
Planetary impact events	678(45)	827 (55)	291(19)	490 (33)	508 (34)	216 (14)	1,505
				Total			
	4,439 (51)	4,253 (49)	1,382 (16)	3,294 (38)	3,034 (35)	982 (11)	8,692

[a] There are no high school students in this category because the question of major field only applies to students in college or graduate programs.

method is also used operationally as a supplemental tool to examine whether anchor items function similarly across different examinee populations on the SAT®. In this study, a two-stage standardization method was used to investigate whether prior content knowledge or cultural background benefits examinees taking TOEFL iBT reading comprehension passages. The standardization approach compares test takers at the same ability level, and the variable used to indicate the ability level is referred to as the matching variable. The total test score is commonly used as a matching variable. In this study, a purification procedure was performed to ensure that the items making up the matching variable were DIF-free (Zenisky, Hambleton & Robin, 2003). The technical details are provided in the following section.

The standardization DIF method. The standardization method, also known as the proportion difference approach, specifies that an item is exhibiting DIF when the expected performance differs on the item for test takers of the same ability level from different groups (Dorans & Holland, 1993). The standardization method examines between-group performance differences after conditioning on some observable variable (e.g., total test score, performance on another similar test). After the test takers from different groups are matched on the target ability, the performance differences on an item may be attributed to factors that are related to group membership. The standardization method focuses on differences in proportion correct, namely the number of test takers who correctly answer an item over the total number of examinees in the focal and reference groups, at each score level (Dorans & Holland, 1993). For the present study, if the maximum score of a reading test is 40, then the test takers are matched on 41 score levels ranging from 0 to 40. The summation of the proportion differences across all score levels indicates the existence of DIF. The standardization approach produces an index of DIF, the standardized *p*-difference ($STD\ P\text{-}DIF$). Mathematically, $STD\ P\text{-}DIF$, or D_{STD} can be defined as:

$$\begin{aligned} STD\ P\text{-}DIF &= \sum_{m=0}^{M} K_{fm}(P_{fm} - P_{rm}) / \sum_{m=0}^{M} K_{fm} \\ &= \sum\nolimits_{m} K_{fm} P_{fm} / \sum_{m=0}^{M} K_{fm} - \sum\nolimits_{m} K_{fm} P_{rm} / \sum_{m=0}^{M} K_{fm}, \\ &= P_f - P_r^* \end{aligned}$$

Where, the subscript *f* refers to the focal group, and *r* refers to the reference group. *m* is the score level and ranges from 0 to the maximum value of the matching variable. K_{fm} is the number of

examinees at score level *m* in the focal group, and $K_{fm} / \sum K_{fm}$ serves as a weighting function in the equation. P_{fm} and P_{rm} refer to the percentage of correct responses at score level *m* in the focal and reference groups, respectively. P_f is the observed performance of the focal group on the item, and P_r^* is the predicted performance of reference group members who are matched with focal group members in ability on the item. In this study, for each passage, examinees with certain content or cultural knowledge are included in the focal group, and the rest of the test-takers responding to that passage are considered members of the reference group. Therefore, positive *STD P-DIF* values suggest that the item potentially advantages focal group members, while negative values suggest that the item disadvantages focal group members.

According to the ETS DIF guidelines (Dorans & Holland, 1993), items can be classified into three DIF categories based on the *STD P-DIF* statistic. *STD P-DIF* values between -.05 and +.05 are considered to demonstrate no or negligible DIF. *STD P-DIF* values between -.10 and -.05 or between +.05 and +.10 are considered to demonstrate moderate DIF. Items displaying moderate DIF may require further inspection. *STD P-DIF* values smaller than -.10 or larger than +.10 suggest substantial DIF, and the items should be carefully examined for the presence of unintended secondary factors (Dorans & Holland, 1993; Penfield & Camilli, 2007). Following the rules for dichotomous items, Dorans and Schmitt (1991, 1993) used a standardized mean difference (*SMD*) index to indicate the DIF effect for polytomous items. *SMD* can be formulated as $SMD = \sum_{m=0}^{M} K_{fm} (\sum_{j=0}^{J} j n_{fjm} - \sum_{j=0}^{J} j n_{rjm}) / \sum_{m=0}^{M} K_{fm}$, where *j* stands for the score category of an item, and n_{fjm} and n_{rjm} are the number of test takers who score *j* on this item with ability level *m* in the focal and reference group, respectively. Let SD_{SMD} denote the pooled standard deviation for the two groups. In this study, *STD P-DIF* was used to examine DIF for dichotomously scored items, and SMD/SD_{SMD} was used for polytomously scored items. Items are classified as having A-DIF if SMD/SD_{SMD} is less than .17 in absolute value *or* not statistically different from zero. Items are classified as having C-DIF if the SMD/SD_{SMD} value is *both* larger than .25 in absolute value *and* statistically different from zero. All other items are classified as having B-DIF (PDIF, 2006).

This method has been used to categorize DIF for polytomous items used in the National Assessment of Educational Progress (NAEP; Allen, Donoghue, & Schoeps, 2001).

The purification procedure. The total reading score of each TOEFL iBT form was used as the matching variable. To ensure the purity of the matching variable, DIF analysis was conducted for all the reading items contained in each form. Items exhibiting B- or C-DIF were removed from the calculation of the total score, and the reduced length total score was then used as the matching variable. This procedure was repeated until no B- or C-DIF items were included in the total score calculation. The purified matching variable was used for the DIF analysis on the key passages. The purification procedure has been proven to be a necessary and effective way to improve the accuracy of DIF identification (French & Maller, 2007; Zenisky, Hambleton, & Robin, 2003, 2004).

Differential Bundle Functioning

A challenge many DIF investigations face is the lack of substantive explanations of DIF results. Theoretical or empirical reasons are needed to speculate about any systematic performance difference between groups of test takers. These explanations may help test developers design items that are less likely to favor any group. Roussos and Stout (1996) proposed a framework for interpreting group differences in performance through bundling items together for differential functioning analysis. This approach employs prior content analysis, likely carried out by content experts, to identify items as bundles that may favor test takers in a particular group on the basis of common item characteristics. The related differential functioning analysis is referred to as differential bundle functioning (DBF) analysis. Compared to the exploratory nature of DIF investigations, DBF analysis is more aligned with confirmatory investigations. It requires a hypothesis that some of the items could be "bundled" for DBF analysis based on certain common characteristics. Many of the more recent studies provide successful examples of applying DBF in explaining occurrences of differential functioning (e.g., Abbott, 2007; Douglas, Roussos, & Stout, 1996; Gierl, 2005; Gierl, Bisanz, Bisanz, & Boughton, 2003; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001). These studies either adopt an a priori approach which involves theoretical speculation as to which items should be grouped for bundle analysis and then confirmation of the hypothesis, or a post-hoc approach, which often involves a content examination of items after the DIF results are available, in order to search for any common characteristics of the DIF items.

This study adopted an a priori approach in constructing the item bundles. An expert panel, composed of the two test developers who helped select the passages, conducted a content analysis of the items in each passage. Items in the three physical science passages were examined for technical terminology. Specifically, the experts examined the extent of the presence of technical or special terms in the item stem and option categories and in the parts of the passages required to answer each of the questions. Although knowing the actual meaning of the terminology may not be required to respond to an item, examinees' familiarity with technical terms may facilitate their performance on reading comprehension, because their familiarity may allow them to predict item difficulty and may increase their level of confidence. Items in the passages about Japanese culture and about European art and the salon were examined for cultural familiarity. Items may contain information familiar to test takers who have experience with these particular cultures. In summary, a number of items on each passage were bundled based on terminology familiarity or cultural familiarity. See Table 1 for the number of items in each bundle in each passage.

To carry out the DBF analysis, a subtotal score was created for the items in each bundle, summing up the individual item scores. The subtotal score can be regarded as a polytomously scored item. For example, if there are five items in the bundle and each item is scored as 0 and 1, then the possible range of scores for this bundle is from 0 to 5. The DBF results tell us whether these items, at an aggregated level, favor one group of examinees or not. The standardization method used for polytomous items was used to conduct the DBF analysis. The matching variable was the same as that used for individual DIF analysis, being the purified total reading score. Finally, in order to examine the effect of content and cultural schemata at the passage level, differential functioning analysis was applied to the entire passage using the sum of the item scores. Similarly, the sum score was treated as a polytomous variable. This analysis is referred to as the differential passage functioning (DPF) analysis. The index SMD/SD_{SMD} was used for both the DBF and DPF analysis.

The software *PDIF* (2006) was used to perform all the DIF, DBF, and DPF analyses. The standardization method used in this study was built in the PDIF program. PDIF is able to detect DIF for both dichotomously and polytomously scored items.

Results and Interpretation

This section presents the results and possible explanations for the identified differential item, bundle, and passage functioning (see Table 3 for results).

European Academic Art and the Salon

For the passage on European academic art and the salon, there were two B-DIF items and one C-DIF item in favor of the focal group (i.e., test takers familiar with European culture). One of the B-DIF items was a reading to learn item. Examinees were asked to complete a brief summary for this passage by selecting three sentences that best expressed the important ideas in the passage. It may be more difficult for examinees who are less well-versed in European culture to understand the gist of this passage. DIF occurrences were also found on similar items in other passages (e.g., restructuring of Japan, landscape painting in Japan) asking test takers to identify the gist or summarize the overall organization of the passage. A plausible explanation is that the focal group members may be more proficient in generating higher level understanding of the passages than examinees who are unfamiliar with these topics.

The C-DIF item was a rhetorical purpose item asking why a particular artist was mentioned in the passage. Rhetorical purpose items measure examinees' ability to identify the author's underlying rhetorical purpose in employing particular expository features in the passage (e.g., to support an argument, to provide an example, to explain a cause). Correct responses require proficiency at inferring the nature of the link between specific features of exposition, in this case mentioning a particular artist, and the author's rhetorical purpose. Familiarity with the artist and with the idea of avant-garde culture in general, which is also mentioned, may help answer this question correctly. The test takers also need to synthesize information from the larger context to identify the right key. It may be easier or faster for European test takers to do this than other test takers because of their specific cultural backgrounds. Therefore, both language characteristics and cultural familiarity may play a role in determining item responses of a particular group. As described earlier, items were bundled together to examine the cumulative effect of differential functioning. The item bundle for this passage, based on European cultural familiarity, showed B-level differential functioning favoring the focal group (Europeans). Therefore, there is some evidence that content schema has an impact on test takers responding to this passage, showing some items favoring examinees who are familiar with the European culture. There was no differential functioning favoring the focal group on the passage level.

Table 3

Results of Differential Item, Bundle, and Passage Functioning

Passages	DIF (*N*)					DBT					DPF				
	A	B		C		A	B		C		A	B		C	
		+	-	+	-		+	-	+	-		+	-	+	-
European art and salon	10	2		1			√				√				
Restructuring of Japan	9	2	2		1	√					√				
Landscape painting in Japan	8	2	1	1	2	√					√				
Planet formation	11	2					√				√				
Planet Venus	12	1				√					√				
Planetary impact events	13		1			√					√				

Note. + indicates that the differential functioning is in favor of the focal group members and – indicates that the differential functioning is in favor of the reference group members. DIF = differential item functioning; DBF = differential bundle functioning; DPF = differential passage functioning. For DBF analysis, since only one bundle was examined for each passage, the √ indicates the magnitude and direction of that bundle. The same is true for the DPF analysis.

Factors Influencing the 13th-Century Restructuring of Japan

One of the passages concerning Japanese culture discussed the factors influencing the 13th-century restructuring of Japan. Two B-DIF items were identified in favor of the focal group (East Asian examinees), and two B-DIF items and one C-DIF item were identified in favor of the reference group (non-East Asians). One of the B-DIF items favoring the reference group was a very difficult item, as judged by the item developers. Getting the correct answer to the item depends on understanding an infrequently used vocabulary word in the passage, possibly suggesting a real vocabulary difference between East Asians and other populations of equal overall ability when understanding of very infrequently used words is required.

The other B-DIF item favoring the reference group required recognizing specific directions in the item stem that asked for the result of an event described in the passage. Again, it seems possible that the non-East-Asian group may be better at inferring rhetorical links, in this case understanding a causal relationship that is inherent but not specifically stated in the passage. One distracter in this item was true and was mentioned in the passage as a result of another event, but it was not the result of the event in question. Additionally, noticing important information in the correct answer required understanding a fair amount of paraphrased information that was not strictly academic vocabulary, again suggesting the non-East-Asians may have a slight advantage over East Asians in mastering infrequently used English vocabulary.

The one C-DIF item favoring the reference group was a difficult sentence simplification item, as judged by the test developers. Sentence simplification items require examinees to select the best simplified version of a complex sentence that retains the essential information of the original. The fact that this item favored the non-Asian group, even though the passage is about East Asian culture, suggests that the non-Asian group may be advantaged over the Asian group for items that include especially difficult lexical and syntactic complexities. One of the B-DIF items favoring the focal group was a negative fact item (i.e., identifying a statement which is not true) concerning changes to agriculture resulting from the restructuring that took place in the 13th century. This item was difficult in that both vocabulary knowledge and information synthesis are needed to respond correctly. This item may have slightly favored East Asians in that this population may favor word matching as a test strategy, and some options in this item could be eliminated using this method. The second B-DIF item favoring the focal group was an

item testing examinees' understanding of the overall organization of the passage. No bundle or passage level differential functioning was identified for this passage.

Monochrome Landscape Painting in Medieval Japan

In this passage, there were two B-DIF items and one C-DIF item favoring the focal group (East Asians) and one B-DIF item and two C-DIF items favoring the reference group (non-East Asians). The B-DIF item favoring the reference group involves figurative and metaphoric language, which may be an area of language proficiency that favors the non-East Asian reference group. One of the C-DIF items favoring the reference group is a relatively difficult item that requires linking an idiomatic phrase in the passage to the correct answer, and the other C-DIF item tests a difficult vocabulary word. Again, the non-East Asians may have a slight vocabulary advantage over East Asians in overall English language abilities because the East Asian languages share little commonality with English, while other languages (e.g., European languages) may be more similar to English. One of the B-DIF items in favor of the focal group was a difficult negative fact item asking which of the four statements about one of the great masters of Japanese landscape painting was *not* true. A similar negative fact item also showed B-DIF in the other passage on restructuring of Japan. The C-DIF item favoring the focal group was a very difficult item that asked examinees to insert a sentence into a place in the passage where this sentence best fits the context. Examinees familiar with this topic may have understood the flow of this passage better and were able to insert the sentence in the proper place. No bundle or passage level differential functioning was identified for this passage.

The Process of Planet Formation

There were two B-DIF items for the passage on planet formation, both favoring the focal group (physical science majors). Both items were relatively difficult as judged by the item developers. One of the items asked examinees to draw an inference from one paragraph in the passage. There were about 10 technical vocabulary words present in the relevant portion of the passage and in the item prompt and option categories. Familiarity with these technical terms may help the focal group members perform better on this item. The second B-DIF item favoring the focal group was a fact item, also with a heavy presence of technical words in both the question and the options. Again, the presence of technical terminology may hinder those who are not familiar with this topic. The item bundle based on technical terminology showed B-level

differential functioning, which suggests that these items overall had an effect on test takers' performance. This finding indicates that for this passage, content schema has a favorable impact on examinees with physical science background. No passage level differential functioning was identified.

The Planet Venus

For the passage on the planet Venus, there was one B-level DIF item favoring the focal group (physical science majors). This item was a vocabulary item with a lot of technical words. This DIF occurrence could possibly be explained by the same reason that the two B-DIF items in the planet formation passage showed B-DIF, that the focal group may be favored by their familiarity with the technical language. Understanding of the general scientific context may also help examinees respond correctly to this item. No bundle or passage differential functioning was identified.

Planetary Impact Events

For the passage on planetary impact events, only one item displayed DIF (B-level), and it favored the reference group, not the focal group. This item was an easy vocabulary item involving no technical terms. It is not clear why this item impeded performance in the focal group. There was no bundle or passage differential functioning.

Discussion

This study investigates the effect of content schemata on TOEFL iBT reading performance. Content schemata consist of prior knowledge or outside knowledge from two sources in this study: (a) systematic training in a particular major field of study and (b) familiarity with a specific culture and its associated art and history. Six passages were investigated for differential item functioning, including three passages on physical sciences and three on cultural topics. Following a confirmatory approach for examining differential functioning, we grouped items together as a bundle within each passage on the basis of presence of culture-related information or technical terminology. The purpose was to examine whether these items as a group had a significant cumulative effect on examinees' reading performance. Each passage was also examined for differential functioning at the passage level.

A general finding is that although these passages heavily involve physical science topics or culture-related content, the great majority of the items displayed no DIF. Additionally, the analysis of many of the items displaying DIF suggests that the differences in performance may be construct-relevant differences based on real differences in certain aspect of the language ability that TOEFL iBT targets. For example, the C-DIF items favoring non-East Asians on the passages concerning Japanese culture provide some evidence that the non-East Asian group have an advantage in understanding difficult English vocabulary over East Asians. And this advantage is associated with a construct-relevant performance difference. The findings support the design principle of the TOEFL iBT that this test is a measure of language skills instead of specific content knowledge. Among the items that showed DIF, the magnitude of the DIF was mainly small to moderate, with very few items displaying large DIF. Only two passages contained a C-DIF item that favored the focal group. The impact of content schema varied across passages, suggesting a possible interaction between prior knowledge and the type of item. For example, consistent DIF was observed favoring the focal groups on summary items and negative fact items. Possibly certain types of items mediate the effect that content schema has on examinees' reading performance.

The majority of the item bundles also showed little or no differential functioning. Two out of the six bundles displayed small to moderate DBF. One of the bundles consists of items about European art and the salon, and the other consists of items about the process of planet formation. The DBF findings suggest that these items, at an aggregated level, favored test takers familiar with the European culture or familiar with the technical terms related to planet formation, which provides evidence of the impact of content schema. Affective factors such as confidence level and anxiety may also play a role in item response behavior. Examinees are more likely to be focused and confident when they encounter passages on topics that they feel comfortable with.

No passage-level differential functioning was identified for any of the passages, suggesting that these passages as a whole did not favor one group or the other. Although a few items and two item bundles displayed DIF, the effect was not strong enough to influence test takers' performance on the entire passage. In addition, the presence of differential item functioning that operated in both directions (some, items favoring the focal group and some favoring the reference group) may lead to the cancellation of some of the passage-level effects.

The DIF items will be closely examined to determine if performance differences represent any bias related to group membership. Distinctions need to be made between DIF occurrence and item bias. Item bias occurs "when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some identifiable characteristic of the test item or testing situation that is not relevant to the test purpose" (Zumbo, 1999, p.12). In this sense, DIF is a necessary but not sufficient condition for item bias. Judgment is required to determine whether the difference in performance is unfairly related to group membership (Zieky, 2003), and whether the difference on an item is fair or not also depends on the purpose of the test and the use of the scores. According to the ETS DIF guidelines (Zieky, 2003), when DIF statistics are available for test items, test developers select questions characterized by A-DIF in preference to items in other DIF categories. In case the number of A-DIF items is not enough for the test to meet all fairness and other specifications, B-DIF items may be used, and items with smaller DIF values are preferred. C-DIF items may not be selected for any test unless these questions are essential to meet important test specifications and the factors that underlie the DIF occurrence are determined not to represent bias (Zieky, 2003). In other words, any items showing C-DIF that are judged to represent bias will not be included in the test.

We also need to keep in mind that DIF may not be a stable item characteristic, in that its occurrence or magnitude may vary across different groups of test takers or across different administrations. McPeek and Wild (1992) reported that the correlation of the DIF index values of items from the same test from two administrations could be as low as .37. Therefore, whether the performance differences indicated by B- or C-DIF represent stable differences awaits further replication and confirmation. In general, passages and items that include technical vocabulary will be carefully scrutinized, made less technical if appropriate, or replaced, since TOEFL is not intended to measure prior knowledge in any field of study. Similarly, passages and items containing culture-specific information will be scrutinized for fairness to examinees from other cultures as part of the passage and item review process.

A potential limitation of this study lies in its classification of members into culture-familiar and culture-unfamiliar groups. This classification was based on test takers' response to the question about which culture they identified themselves as most familiar with. There may be an interaction between cultural familiarity and personal interest. For example, people who grow

up in European countries may be very interested in Asian culture and be knowledgeable about it. In this case, they should be grouped into the culture-familiar group on items concerning Asian culture even though they identify themselves as most familiar with European culture. This hypothesis should be tested in future studies examining the impact of cultural knowledge on TOEFL iBT reading performance.

For future studies clarifying the effect of content schema on TOEFL iBT reading performance, examinees' language proficiency should be further examined. The DIF and DBF analyses in this study were conducted using test takers of all proficiency levels. There may be an interaction between an examinee's major field/cultural knowledge and language proficiency level, which may change the magnitude or even direction of the DIF or DBF occurrences. For example, when test takers of high English language proficiency are separated for analysis, we may not be able to observe any impact of content schema, because this group of examinees is at such a high level of proficiency that whether or not they have prior knowledge does not interfere with their test performance. Therefore, it may be the low to intermediate level test takers who are mostly affected by content schema.

References

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing, 24,* 7-36.

Abu-Rabia, S. (1996). Factors affecting the learning of English as a second language in Israel. *The Journal of Social Psychology, 136*, 589-595.

Alderson, J. C., & Urquhart, A. H. (1984, November). *ESP tests: The problem of student background discipline.* Paper presented at the international symposium on language testing, Tampere, Finland.

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (Eds.). (2001). *The NAEP 1998 technical report* (NCES 2001-509). Washington, DC: National Center for Education Statistics.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, England: Cambridge University Press.

Brown, J. D. (1982). Testing EFL reading comprehension in engineering English (Doctoral dissertation, University of California, Los Angeles). *Dissertation Abstracts International, 43*, 1129A-1130A.

Chihara, T., Sakurai, T., & Oller Jr., J. (1989). Background and culture as factors in EFL reading comprehension. *Language Testing, 6,* 143-151.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1997: An application of the standardization approach* (ETS Research Rep. No. RR-83-09). Princeton, NJ: ETS.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Rep. No. RR-91-47). Princeton, NJ: ETS.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R.E. Bennett & W.C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 135-165). Hillsdale, NJ: Lawrence Erlbaum Associates.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*(4), 465-484.

Enright, M., Grabe, W., Koda, K., Monsethal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph Series Rep. No. 17). Princeton: NJ: ETS.

Erickson, M., & Molloy, J. (1983). ESP test development for engineering students. In J. Oller, (Ed.), *Issues in language testing research* (pp. 280-288). Rowley, MA: Newbury House.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*(2), 113-148.

Floyd, P., & Carrell, P. L. (1987). Effects on ESL reading of teaching content schemata. *Language Learning, 37*, 89-108.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*(3), 373-393.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24*(1), 3-14.

Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. *Journal of Educational Measurement, 40*, 281-306.

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20,* 26-36.

Hale, G. (1988). *The interaction of student major-field group and text content in TOEFL reading comprehension* (TOEFL Research Rep. No. RR-25). Princeton, NJ: ETS.

Holland, P. W. (1985, October). *On the study of differential item performance without IRT.* Paper presented at the meeting of the Military Testing Association, San Diego.

Holland, P. W., & Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland P. W. & Wainer, H., (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Keshavarz, M. H., Atai, M. R., & Ahmadi, H. (2007). Content schemata, linguistic simplification, and EFL readers' comprehension and recall. *Reading in a Foreign Language, 19,* 19-33.

McPeek, W. M., & Wild, C.L. (1992). *Identifying differentially functioning items in the NTE core battery* (ETS Research Rep. No. RR-92-62). Princeton, NJ: ETS.

PDIF: Compute and print DIF statistics [Computer program from F4STAT for Fortran 90 statistical subroutine library]. (2006). Princeton, NJ: ETS.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125-167). New York: Elsevier.

Roussos, L., & Stout, W. (1996): A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355–371.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. E. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing, 17*, 85-114.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement, 63*(1), 51-64.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9*(1-2), 61-78.

Zieky, M. (2003). *A DIF primer*. Princeton, NJ: ETS.

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, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendix A

E-Mail to the Test Takers

Dear TOEFL iBT test taker:

ETS is currently conducting a research study on the Reading section of the TOEFL iBT test. As part of this research study, ETS is inviting previous TOEFL iBT test takers to complete a short online survey. The survey primarily seeks to gain information on the respondent's educational background. Information shared on the survey will be used for research purposes only and has no bearing on the respondent's past or future test score/s. The survey does not require the respondent to share any identifying information such as name, TOEFL test ID, etc.

By completing the survey, respondents can receive one of many cash prizes:

The first 100 respondents will each receive a check worth **$75**

The next 200 respondents will each receive a gift voucher worth **$50**

As a previous TOEFL iBT test taker, you are invited by ETS to complete the online survey and possibly win a prize!

Click on the link below (or copy and paste the link into your browser window) to complete your online survey. You are requested to complete the survey based on your …… admin date.

(Insert the URL here)

We thank you for your time and wish you the best in the future!

ETS
Princeton, NJ

Appendix B

Background Survey for TOEFL iBT

Dear Respondent,

You took the TOEFL iBT test on …… You are requested to answer the questions in the survey from the point of view of this test administration date.

Please answer all of the questions as accurately as possible.

Thank you for your time.

---

1. Please state your native language
   (blank) (required response)

   _____

2. Gender
   (drop-down menu) (required response)
   - ❑ Female
   - ❑ Male

3. In which region did you grow up?
   (drop-down menu) (required response)
   - ❑ East Asia
   - ❑ South Asia
   - ❑ South-east Asia
   - ❑ Middle-East
   - ❑ Scandinavia
   - ❑ Eastern Europe
   - ❑ Western Europe
   - ❑ Africa
   - ❑ North America
   - ❑ South America
   - ❑ Australia/New Zealand
   - ❑ Other (please specify)

     _____

4. Which culture do you most identify with?
   (drop-down menu) (required response)
   - ❑ East Asian Culture
   - ❑ South Asian Culture
   - ❑ Middle-Eastern Culture
   - ❑ Scandinavian Culture
   - ❑ Eastern European Culture
   - ❑ Western European Culture
   - ❑ African Culture
   - ❑ North American Culture
   - ❑ South American Culture
   - ❑ Australian/New Zealand Culture
   - ❑ Other (please specify)

     _____

5. From the topics listed, select the one that you are most interested in.
   (drop-down menu) (required response)
   - ❑ Botany
   - ❑ European Art
   - ❑ Japanese Culture
   - ❑ East Asian Art
   - ❑ Planetary Sciences
   - ❑ None of the above (If this option is selected then "Other" required)
   - ❑ Other (please specify)

     _____

6. What was your academic status when you took the TOEFL iBT test?
   (drop-down menu) (required response)
   - ❑ Secondary/High school student
   - ❑ Undergraduate/College student
   - ❑ Graduate student
   - ❑ Other (Please specify) _____

7. If you were a college or graduate student when you took the TOEFL iBT, what was your major field of study?
   (drop-down menu) (in case of 'NA', skip to Q.8)
   - ❑ (the department codes in the TOEFL bulletin were used)
   - ❑ Not applicable

8. If you were a college or graduate student when you took the TOEFL iBT, select the option that describes your status at the time of your TOEFL iBT test.
   (drop-down menu) (required response)
   - ❑ I had just started my program in my department
   - ❑ I was half-way through my program in my department
   - ❑ I had almost finished my program in my department

9. Have you taken the TOEFL iBT more than once?
   (drop-down menu) (required response)
   - ❑ Yes
   - ❑ No (Skip Q.12)

10. If yes, please specify how many times you have taken the TOEFL iBT
    (drop-down menu) (Required response)
    - ❑ Twice
    - ❑ Three times
    - ❑ Four times or more

Please provide a name and an address for us to mail you your check.

Name _____

Address (line 1) _____

Address (line 2) _____

City _____

State/Province & Zip _____

Country _____

You have reached the end of the survey. Click on **Submit Responses** to post your responses and exit the survey. Thank you for your time!

❑ Submit Responses

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

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

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

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

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

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