ETS TOEFL

Quality Beyond Measure.

TOEFL iBT® Research Report
TOEFL iBT–18

Is There Any Interaction Between Background Knowledge and Language Proficiency That Affects *TOEFL iBT®* Reading Performance?

Yao Zhang Hill

Ou Lydia Liu

October 2012

Is There Any Interaction Between Background Knowledge and Language Proficiency That Affects *TOEFL iBT*® Reading Performance?

Yao Zhang Hill
Kapi‘olani Community College, University of Hawaii

Ou Lydia Liu
ETS, Princeton, New Jersey

RR-12-22

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

This study investigated the effect of the interaction between test takers' background knowledge and language proficiency on their performance on the *TOEFL iBT*® reading section. Test takers with the target content background knowledge (the focal groups) and those without (the reference groups) were identified for each of the 5 selected passages based on their self-identified academic and cultural backgrounds. The test takers were further classified into high and low proficiency groups based on their TOEFL iBT scores. Differential functioning was investigated at the item, item bundle, and passage levels. The results suggested that background knowledge interacted with language proficiency on certain items, which could be attributed to idiosyncratic passage and item characteristics (i.e., characteristics that were specific to a particular passage or item). Only 1 of the 5 passages investigated showed intermediate differential bundle functioning, favoring the focal group for both the high and low proficiency groups. There was no differential functioning at the passage level. This research sheds new light on our understanding of the effects of background knowledge and its interaction with language proficiency in the context of second language reading comprehension. It also has significant practical implications for test developers in advancing fair assessments.

Key words: background knowledge, differential item functioning, differential bundle functioning, TOEFL iBT

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 (2011-2012) members of the TOEFL Committee of Examiners are:

John M. Norris (Chair)	University of Hawaii at Manoa
Fanta Aw	American University
Barbara Hoekje	Drexel University
Ari Huhta	University of Jyväskylä
Eunice Eunhee Jang	University of Toronto
John M. Norris	University of Hawaii at Manoa
James Purpura	Columbia University
John Read	University of Auckland
Carsten Roever	University of Melbourne
Steve Ross	University of Maryland
Norbert Schmitt	University of Nottingham
Robert Schoonen	University of Amsterdam
Ling Shi	University of British Columbia

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

Table of Contents

List of Tables

Page

Reading comprehension is a complex mental activity and is believed to be affected by many factors and their interplay. Among the different factors that influence reading comprehension, background knowledge—and content background knowledge in particular—has been extensively studied (see a review by Hudson, 2007). Interest in the effect of content background knowledge was first seen in the work of top-down comprehension theorists (Goodman, 1976; Smith, 1994). The reading process in the top-down model was viewed as what Goodman (1976) called a psychological guessing game; he proposed that comprehension heavily relies on the reader's existing syntax, semantic, and experiential knowledge, and that this comprehension requires minimum language cues. Contrasted with the top-down model of comprehension is the bottom-up model. The latter suggests that comprehension follows a linear serial processing from decoding all graphemic and phonemic information of a word before understanding its meaning. In other words, the model assumes that a reader needs to visually process each letter of the word and pronounce the word in the brain before understanding the meaning of a word. The model also assumes that understanding at higher levels (intrasentential, intersentential, interparagraph, and global understanding) cannot be achieved without the low level processing of each and every linguistic cue. Most scholars today would accept an interactive approach to reading, believing that comprehension involves both bottom-up and top-down processing. Reading comprehension is a parallel process of the low-level linguistic association net (text-base) and the global-level world knowledge association net (knowledge-base) with nonlinear interactions (Kintsch, 1988) along with constant shaping of each other's association net. It places as much emphasis on background knowledge as on lower level linguistic processing.

Empirical investigations have shown that the effect of background knowledge is very complex and interacts with many text, task, and reader variables (Alderson, 2000). Among these variables, language proficiency has been found to be an important factor that interacts with background knowledge in reading comprehension tasks (Chan, 2003; Chen & Donin, 1997; Clapham, 1996; Hammadou, 1991; Hock, 1990; Usó-Juan, 2006).

To understand the effect of background knowledge in large scale language testing is extremely important because, on the one hand, we want test takers to utilize their background world knowledge (Alderson, 2000); but on the other hand, we do not want to jeopardize test fairness by providing passages that give inadvertent advantages or disadvantages to students from a specific academic or cultural background.

The focus of the current study is to investigate the effects of the interaction between background knowledge and language proficiency on a reading comprehension test at the item level. The test being investigated is the *TOEFL iBT*® exam. TOEFL iBT was introduced in September 2005 to replace the TOEFL computer-based test (CBT). There were substantial changes in both test format and content in the reading section of TOEFL iBT when compared to CBT, and TOEFL iBT now consists of fewer but longer reading texts (about 700 words each) than CBT passages (250–300 words each). The number of items has also increased from 11 to 14 per passage. In addition to the changes in test format, the test content of TOEFL iBT has been extended to include materials regarding the people, culture, and history of places other than the United States and Canada, which was rarely the case on the CBT.

The reduced number of passages leads to a reduction in topic variety on the TOEFL iBT. One concern is that the reduced topic variety and the introduction of new cultural topics may advantage test takers from particular academic or cultural backgrounds. Liu, Schedl, and Kong (2009) investigated whether major field of study and cultural background had an impact on test takers' TOEFL iBT performance and concluded that no consistent impact was identified. However, Liu et al. did not consider the role of language proficiency in their investigation. To further advance the research in this line of inquiry, we reanalyzed the data used in the Liu et al. study to address the interaction between background knowledge and language proficiency.

Literature Review

In this review, we focus on two types of content background knowledge: discipline domain-specific (DDS) knowledge and cultural familiarity. We then look at their effect on reading comprehension and the interaction effect between background knowledge and language proficiency.

Discipline Domain-Specific (DDS) Background Knowledge and Language Proficiency

The question of what role DDS background knowledge plays in academic reading has interested many researchers in reading comprehension and practitioners in the field of language testing (Alderson & Urquhart, 1985a, 1985b; Brown, 1984; Chen & Donin, 1997; Clapham, 1996; Erickson & Molloy, 1983; Hale, 1988; Hock, 1990; Park, 2004; Peretz & Shoham, 1990; Salager-Meyeer, 1994; Usó-Juan, 2006). Very often researchers have found the effect of DDS knowledge to be significant, and when such an effect has been identified, it has been found to facilitate reading comprehension (Alderson & Urquhart, 1985b; Chen & Donin, 1997; Clapham, 1996; Hale, 1988; Hock, 1990; Peretz & Shoham, 1990; Salager-Meyeer, 1994; Usó-Juan, 2006). Many researchers have also directly investigated the effect of language proficiency, or linguistic competence, along with the effect of background knowledge on reading comprehension (Brown, 1984; Chen & Donin, 1997; Clapham, 1996; Erickson & Molloy, 1983; Hock, 1990; Park, 2004; Salager-Meyeer, 1994; Usó-Juan, 2006), or indirectly examined the effect of language proficiency when they studied the effect of background knowledge on academic reading comprehension (e.g., Alderson & Urquhart, 1985b; Peretz & Shoham, 1990). Among these studies, the ones that had relatively large sample sizes (e.g., Brown, 1984; Clapham, 1996; Erickson & Molloy, 1983; Hock, 1990; Park, 2004; Usó-Juan, 2006) found that language proficiency exerted a much greater effect than DDS background knowledge on reading comprehension. For example, Clapham (1996) conducted multiple regressions using different indicators of DDS background knowledge and grammar test scores as the English language proficiency indicator to predict reading comprehension scores for 842 IELTS test takers. The regression model explained 45% of the variance in the reading scores when all the texts were included in the model. Among all the predictors, proficiency alone explained 44% of the variance. When only texts with a high level of subject area specificity were included in the regression model, the model explained a total of 38% of the variance in reading scores and proficiency alone explained 26% of the variance. Background knowledge variables only accounted for 1% and 12% of the variance, respectively, when predicting scores for all reading

texts and for subject area specific texts.

Hock (1990) found that by employing multiple regression for a total of 317 Malaysian undergraduate students (with 141 majoring in medicine, 95 in law, and 81 in economics), DDS background knowledge and English language proficiency both significantly predicted reading comprehension scores on the texts in the test takers' own subject areas, and both variables combined accounted for 50%, 62%, and 46% of the variance of the scores on texts in medicine, law, and economics, respectively. The standardized regression coefficients of language proficiency in all regression models weighed about twice as much as that of DDS background knowledge in predicting the reading comprehension scores of the texts in test takers' own subject fields, indicating a stronger effect of language proficiency compared to background knowledge. Language proficiency became the only significant variable when predicting scores on the texts outside of test takers' subject areas in three out of four regression analyses.

Usó-Juan's (2006) study produced similar results to those produced by Hock (1990). She found that for 360 Spanish-speaking undergraduate students, DDS background knowledge and language proficiency both significantly predicted reading scores on six English-for-academic-purpose texts through six simple regressions, and the two predictors combined accounted for 41.3% to 47.9% of the reading score variance. The standardized regression coefficients of English knowledge weighted 1.83 to 3 times as much as that of background knowledge, implying a stronger effect of proficiency compared with that of background knowledge.

When the interaction effect between language proficiency and DDS background knowledge is examined, the findings are not consistent. Some studies have reported no interaction effect between proficiency and background knowledge (e.g., Brown, 1984; Erickson & Molloy, 1983), but others have found an interaction effect (e.g., Alderson & Urquhart, 1985b; Clapham, 1996; Koh, 1985).

Both Brown (1984) and Erickson and Molloy (1983) investigated the effect of background knowledge and language proficiency on an engineering reading comprehension test given to both engineering and nonengineering college students. There were two proficiency

groups: native speakers of English and nonnative speakers of English. The results showed that engineering majors performed better than nonengineering students, and native speakers scored higher than nonnative speakers. However, the interaction between language proficiency and background knowledge was not significant.

On the contrary, Clapham (1996) found that the effect of background knowledge, judged by subject area of study, was different on the IELTS academic reading comprehension texts across three proficiency levels as determined by scores on a grammar test. On two of the three revised academic modules that consisted of texts with higher content specificity, the effect of background knowledge was not evident on reading scores for the students with grammar scores below 60%, but the background knowledge effect was significant in all three modules for those who scored 80% and above. The results from these content-specific text modules, as the author claims, "seem[ed] to confirm that L2 readers make more use of background knowledge as their language proficiency improves" (p. 184).

To complicate matters further, in Koh's (1985) study, 60 college science students in Singapore with higher English proficiency levels scored similarly on the texts in their area of study and the unfamiliar texts on politics and history, showing only a slight advantage of background knowledge. However, for another 60 science students with lower English levels, the effect of background knowledge was more evident, in that they scored much higher on the familiar science text than on unfamiliar texts on politics and history. Since the author did not provide information on the standard deviations, effect sizes could not be calculated, but the descriptive statistics were indicative of an interaction effect between proficiency and background knowledge.

Cultural Background Knowledge and Language Proficiency

Studies investigating the interaction between cultural knowledge and language proficiency have reported mixed findings as well (Chan, 2003; Johnson, 1981; Keshavarz, Mahmoud, & Ahmadi, 2007).

Johnson (1981) gave a culturally familiar and unfamiliar English text to 46 Iranian

intermediate/advanced English as second language (ESL) students and 19 American college students. Cultural familiarity had no significant effect for American readers, but its effect was significant for Iranian readers, who recalled the story from their own cultural origin better. The results pointed to an interaction effect between cultural familiarity and language proficiency.

Chan (2003) investigated the effect of cultural content and reading proficiency with learners of English from mainland China and Hong Kong (HK). Students were asked to take cloze tests made from a general content text on sleep and a text based on HK subculture entitled *Symbols of Hong Kong*. Two two-way ANOVA (regional background x proficiency level) analyses were conducted on scores on the cloze tests made from each of the two texts. Both ANOVA results showed that language proficiency had a consistent significant effect across the two texts. The ANOVA analysis with HK passage cloze test scores as the dependent variable revealed a significant interaction effect between proficiency and background knowledge. The post-hoc pair-wise comparisons showed that the effect of background knowledge was more evident in lower proficiency students. Lower proficiency HK and mainland Chinese students did not score differently on the sleep cloze test, but lower proficiency HK students scored significantly higher than the mainland Chinese counterparts on the cloze tests with the HK culture content. In contrast, there was no score difference between higher proficiency HK and mainland Chinese students on either test.

Keshavarz, Mahmoud, and Ahmadi (2007) investigated three variables that potentially affect reading comprehension: cultural background knowledge, linguistic simplification of the text, and language proficiency. In their study, 240 Iranian English as foreign language (EFL) students were divided into four groups, each having two proficiency levels. Each group was assigned to read a culturally familiar text on their religious leader and a culturally unfamiliar text. Two ANOVA analyses, with dependent variables being multiple-choice scores for one analysis and the recall scores for the other, revealed that cultural background knowledge had a strong facilitative effect on comprehension—so did language proficiency, but not their interaction. In this study, the effect of cultural background knowledge had a consistent effect on

comprehension regardless of the readers' proficiency level.

To sum up, in general, language proficiency appears to be a stronger predictor of academic reading performance than background knowledge in general. There are times when language proficiency interacts with content background knowledge in affecting reading comprehension. Reasons for the inconsistent findings of the interaction effect likely vary, but one important reason may be the use of nonstandardized measures of language or linguistic proficiency. Since researchers have employed different instruments to test language proficiency, definitions of the high proficiency group in one study may differ from that of other studies.

Differential Item Functioning (DIF) Analysis of Background Knowledge and Language Proficiency

All the aforementioned studies on background knowledge and its interaction with proficiency levels were carried out using t-test, ANOVA, and regression analyses, and sometimes only descriptive statistics were reported. Test takers' item-level performance was often ignored, and item type and item difficulty were not controlled. Therefore, while it is possible that students from a particular background may have performed better on certain items but not on others, use of the total score may have cancelled out the effect of background knowledge on individual items. For test developers, understanding the behavior of individual items is very important for identifying potential item bias and reducing or eliminating biased items. Investigations into item-level difference between high- and low-knowledge groups can motivate reading comprehension theories to explore reading process under different reading conditions and explore the dynamic interaction between text, task, and reader characteristics (Alderson, 2000).

Differential item functioning (DIF) analysis is usually used to investigate whether performance differs between different racial, ethnic, and gender groups at the item level for test takers with matching ability or proficiency levels. DIF occurs when two groups of test takers of the same ability level show a different probability in answering an item correctly (Camilli & Shepard, 1994). The occurrence of DIF indicates that the test may have a secondary dimension in addition to the primary dimension that assesses the construct-related knowledge or skills

(Roussos & Stout, 1996). If the secondary dimension is construct-irrelevant, then DIF items are potential threats to test validity. DIF analysis has been a common method in detecting item bias to ensure test fairness. However, the occurrence of DIF does not automatically entail the existence of item bias, as the second dimension uncovered through DIF analysis may be part of the construct intended by the test. DIF is a necessary but not sufficient condition for item bias. DIF only manifests statistical differential performance on an item for two groups of the same ability, but whether the items can be considered biased depends on the evaluative process, usually through content analysis of test task, to determine whether such difference reflects construct-relevant or irrelevant variance. For example, Chen and Henning (1985) found that all four vocabulary items on UCLA's ESL placement test, which favored Spanish speakers, tested English words containing Spanish cognates. The vocabulary test reflected the fact that many English words have Spanish cognates. Therefore knowing these words constituted construct-relevant knowledge. In this case, DIF did not indicate item bias. Nevertheless, the detection of DIF should be the first warning sign for the test developers and warrants close examination of the fairness of DIF items.

In the literature on second language DIF analysis, the commonly investigated group variables include native language (Abbott, 2007; Elder, McNamara, & Congdon, 2003; Jang & Roussos, 2009; Kim & Jang, 2009; Lee, Breland, & Muraki, 2005; Ryan & Bachman, 1992; Sasaki, 1991), gender (e.g., Breland, Lee, Najarian, & Muraki, 2004; Roznowski & Reith, 1999; Ryan & Bachman, 1992; Singh, 2005; Takala & Kaftandjieva, 2000) and ethnicity (e.g., Roznowski & Reith, 1999; Snetzler & Qualls, 2000). (See reviews on DIF in language assessment in Ferne & Rupp, 2007, and McNamara & Roever, 2006.) Although some of these studies have tried to explain the occurrence of DIF with either a focal or reference group's background knowledge, the background knowledge referred to in the explanation is often linguistic background (such as Indo-European speakers knowing more English vocabulary because of their familiarity with English cognates compared to their non-Indo-European counterparts, as in Jang & Roussos, 2009), language training, or test preparation background.

(For example, Chinese students performed better on grammar items in several studies, which was explained by their intensive training in English syntax, as in Sasaki, 1991.) Only a few studies have specifically investigated academic background (e.g., Pae, 2004) or cultural background (e.g., Banks, 2006; Davidson, 2004), and even fewer studies have looked at the interaction between content background knowledge and language proficiency (Davidson, 2004).

There is evidence that students with a certain cultural or academic background have advantages on items that contain content related to that background when compared to the students of other group membership with equal ability. For example, Banks (2006) found that Hispanic and African-American students had a higher probability of answering multiple-choice items correctly on the reading and language arts subtests on the Terra Nova test compared to the white students when the content of those options reflected their cultural values. On the English subtest of the Korean National Entrance Exam to Colleges/Universities, Pae (2004) found seven items that favored students majoring in humanities were about human relations. Nine items that favored students in scientific fields dealt with science-related topics, data analysis, and number counting.

Studies that have investigated DIF with regard to native language or ethnicity groups have often focused on the influence of cultural background on test performance, with consideration given to linguistic background (Chen & Henning, 1985; Kim & Jang, 2009; Sasaki, 1991), language processing strategies (Abbott, 2007), test-taking purpose (Ryan & Bachman, 1992), and other sociocultural factors that may also contribute to the difference in performance. Roever (2007) reported that European test takers, mostly German speakers whose culture was grounded in Judeo-Christian tradition, scored higher on the pragmatic question “Is the Pope Catholic?” compared to their Asian counterparts, while the Asian group showed equal or better performance in answering another question, “Do fish swim?”, which had the same pragmatic meaning—the answer to the question is obviously yes. The likely explanation of the weaker performance on the Pope question for the Asian students is that they lacked Judeo-Christian cultural knowledge. Kim and Jang (2009) found that recent immigrants to Canada

performed less well on an item that required students to define a cultural symbol of Canada than their native-speaker counterparts on the Ontario Secondary School Literacy Test given to 10th grade students.

However, the influence of cultural background is not always that clear. Davidson (2004) investigated the effect of cultural difference between aboriginal and nonaboriginal Grade 4 and Grade 7 students on the Grades 4 and 7 Reading and Numeracy Foundation Skills Assessment of the British Columbia Ministry of Education. Using the items consistently identified as employing both Linn-Harnisch (LH) and logistics regression detection methods, the author found that the majority of the reading items favored the aboriginal students. However, the only item with large differential functioning magnitude detected using the LH method favored the nonaboriginal students. The author was unable to offer plausible explanations for the sources of DIF.

A common issue facing exploratory DIF investigations is the lack of meaningful explanation for the DIF that is detected. Very often, when researchers set out to examine all the items on a test without a hypothesis as to whether a certain item is likely to exhibit DIF, the study ends up with the sources of DIF left unexplained (e.g., Snetzler & Qualls, 2000), failure to provide a plausible explanation of the sources of DIF (e.g., Davidson, 2004) despite an effortful search for it, or simply attributing DIF to the diversity of the population (e.g., Elder et al., 2003). In recognition of such problems, several studies have used both exploratory and confirmatory approaches in DIF detection through differential bundle functioning (DBF) analysis (e.g., Abbott, 2007; Banks, 2006) or differential test functioning (DTF) analysis (e.g., Pae & Park, 2006). The idea behind DBF is to have content experts identify the items that may exhibit DIF according to some criteria, such as cultural values (Banks, 2006) or types of reading strategies (Abbott, 2007), and then to analyze these items together as if they were one polytomously scored item. Since in DBF analysis the items have already been classified based on content differences, when differential functioning is detected, the source can be reasonably attributed to the original criterion used for the classification of the item bundles, thus confirming whether the hypothesis regarding the sources of DIF can be supported. For DTF analysis, the whole test is treated as one

item, which is then subject to DIF analysis.

To the best of our knowledge, the only study that has explicitly investigated the effect of academic and cultural background knowledge on reading comprehension, using both exploratory and confirmatory DIF detection methods, is that of Liu et al. (2009). They investigated DIF, DBF, and differential passage functioning (DPF) for six TOEFL iBT reading passages (n = 8,692): three about physical science and three containing non-American cultural content. The majority of the items displayed no DIF. Two out of six bundles were found to favor the physical science examinees or examinees familiar with a certain culture. Passage-level differential functioning was not found. The conclusion of Liu et al. was that the TOEFL iBT test takers' background knowledge had a minimal effect on their performance in reading comprehension.

As informative as the Liu et al. (2009) study was, it did not consider the interaction between background knowledge and language proficiency, and therefore, the results may not hold when proficiency is included in the investigation. To contribute to the ongoing research on the validity of TOEFL iBT scores, we examined the following research question in this study: Does the effect of content background knowledge on the TOEFL iBT reading comprehension test differ across test takers' proficiency levels?

Methods

Instrument

The data used in this study came from the Liu et al. (2009) study, in which six passages were selected for analysis. Since the current study considers both background knowledge (i.e., major field of study or cultural familiarity) and language proficiency, it was crucial to make sure that the sample size was adequate for both the high- and low-proficiency groups examined for DIF analysis. One of the six passages on physical science was excluded from this study due to its small sample size (there were fewer than 60 examinees who were physical science majors in the low language proficiency category). The remaining five passages came from four TOEFL iBT tests administered between October 2007 and February 2008, which had a large number of test takers (n > 10,000). The length of these five passages ranged from 678 to 715 words. All the

passages were relatively difficult as judged by the Flesch Reading Ease index (Flesch, 1948), ranging from 23.8 to 39.9 out of a total range from 0 to 120, with the lower numbers indicating a higher degree of difficulty. Texts with a Flesch Reading Ease index from 0 to 30 are best understood by university students (Flesch, 1948).

Three passages were on cultural topics: European art history, the restructuring of Japanese society in history, and the evolution of a style of painting in Japan. The remaining two passages were about physical science: planet formation and how collision and planetary impact shapes planets. These five passages will be referred to as European Art (P1), Restructuring of Japan (P2), Painting in Japan (P3), Planet Formation (P4), and Planetary Impact (P5) in this paper, where P3 and P5 were on the same TOEFL iBT test.

Each passage contained 13 multiple-choice, dichotomously scored items and one multiple-selection, polytomously scored item. The point range of the polytomously scored item was from 0 to 2. Therefore, the maximum possible score for each passage was 15. For the purposes of this paper, all the items on one passage will be referred to as a testlet. The reliability of each testlet was reasonably high (above .70). The passage characteristics for all the investigated testlets are presented in Table 1.

Table 1

Passage Characteristics

Passage Each with 14 items	Number of words	Flesch Reading Ease scale: 0–120	Reliability Cronbach's alpha
1. European Art (P1)	714	23.8	.77
2. Restructuring of Japan (P2)	711	28.3	.71
3. Painting in Japan (P3)	678	39.9	.70
4. Planet Formation (P4)	695	29.2	.77
5. Planetary Impact (P5)	715	30.0	.72

Test Takers

We obtained background information and test scores from 7,310 TOEFL iBT test takers. These test takers provided background information through an email survey, and their test scores were obtained from ETS. We classified students into two proficiency levels based on their TOEFL iBT total scores. Proficiency cutoff scores were determined by examining the total score distributions for each of the four TOEFL iBT tests that contained the target passages and selecting a score close to the median on all four distributions. All four score distributions were negatively skewed. The total possible score on the TOEFL iBT ranges from 0 to 120, and the majority of the test takers scored above 90. The medians on the four tests were: 97, 92, 95, and 92, for tests containing P1, P2, P3 and P5, and P4. Although 90 was slightly lower than the median on four distributions, 0 to 90 theoretically already covers a wide range of proficiency. Therefore, we chose to use 90 as our proficiency cutoff score. Test takers who scored between 0 and 90 were classified as the lower proficiency group, and those scoring between 91 and 120 were classified as the higher proficiency group. Proportionally, there were 37%, 44%, 43%, and 46% of test takers in the lower proficiency group for the TOEFL iBT tests that contained P1, P2, P3 and P5, and P4.In other words, lower proficiency students accounted for a smaller proportion of test takers on those four tests, despite the fact that there was a theoretically wider range of proficiency (0–90) among the lower proficiency group.

There are other ways to define two proficiency groups, such as excluding test takers with midrange scores or only including two groups of test takers with mean test scores of nonlapping confidence intervals. However, these methods of classification would have significantly reduced the number of test takers that could be used for the DIF study. Demographic information on the test takers is summarized in Table 2.

We also classified students into focal and reference groups for DIF analysis on each testlet. We followed the same focal and reference group classification procedure as used in Liu et al. (2009, p. 7). Focal groups consisted of students who had background knowledge related to the content of the passage. For example, the focal group for the European Art (P1) passage consisted

Table 2

Test Takers' Demographic Information

	Focal group					Reference group					
Passage	Male *n* (%)	College students *n* (%)	Graduate students *n* (%)	High school students *n* (%)	Total[a] *N*	Male *n* (%)	College students *n* (%)	Graduate students *n* (%)	High school students *n* (%)	Total *N*	Grand total *N*
1. European Art (P1)	42(39)	47(44)	35(33)	11(10)	107	362(57)	297(47)	227(36)	56(9)	633	740
2. Restructuring of Japan (P2)	130(47)	110(39)	59(21)	82(29)	279	317(49)	223(34)	252(39)	107(17)	647	926
3. Painting in Japan (P3)	87(42)	85(41)	52(25)	51(25)	208	288(50)	198(34)	230(40)	83(14)	579	787
4. Planet Formation (P4)	80(67)	60(50)	59(50)	0	119	234(45)	168(32)	174(33)	106(20)	525	644
5. Planetary Impact (P5)	99(65)	74(49)	78(51)	0	152	276(43)	209(33)	204(32)	134(21)	635	787
1. European Art (P1)	155(47)	144(44)	106(32)	38(12)	330	510(55)	447(48)	283(31)	101(11)	925	1,255
2. Restructuring of Japan (P2)	72(50)	33(23)	49(34)	53(37)	145	517(51)	373(37)	357(35)	149(15)	1013	1,158
3. Painting in Japan (P3)	34(34)	23(23)	36(36)	28(28)	99	514(54)	342(36)	371(39)	127(13)	952	1,051
4. Planet Formation (P4)	112(76)	80(54)	67(46)	0	147	304(50)	195(32)	196(33)	107(18)	602	749
5. Planetary Impact (P5)	154(79)	88(45)	107(55)	0	195	394(46)	277(32)	300(35)	155(18)	856	1,051
Grand total[b]	712	582	463	263	1,434	3,046	2,243	2,090	836	5,876	7,310

[a]The total number of test takers for the focal and reference groups is the sum of high school, college, and graduate students, as well as test takers with missing data, in those groups, respectively. [b]Because the two passages Painting in Japan (P3) and Planetary Impact (P5) are from the same test, the grand total is the unique total number of test takers on these five passages.

of students who felt most closely identified with Eastern European, Western European, or Scandinavian culture. The reference group included everyone else who took that testlet. See Table 3 for the sample size of each group by proficiency level. Traditionally, a focal group consists of test takers who are potentially disadvantaged and who are the target of the investigation. In this study, students with assumed background knowledge were the primary group of interest for investigation, and therefore they were assigned to the focal group. We set the minimum sample size at 100 for each focal and reference subgroup, but still included the passage Restructuring of Japan (P2), which had 99 examinees in its focal/high proficiency group. Fidalgo, Ferreres, and Muñiz (2004) showed that when the minimum sample size for both the reference and focal groups was set at 100, the standardization method used to detect DIF resulted in smaller Type I error rates (.12 and .13 for detecting moderate and large DIF, respectively) than for other sample size combinations where the minimum sample size for the focal group was 50. For this reason, we use 100 as the minimum sample size for our focal group.

Table 3

Number of Test Takers for Differential Functioning Analysis Conducted on Each Testlet

	Low proficiency		High proficiency	
Passage	Focal	Reference	Focal	Reference
1. European Art (P1)	107	633	330	925
2. Restructuring of Japan (P2)	279	647	145	1,013
3. Painting in Japan (P3)	208	579	99	952
4. Planet Formation (P4)	152	635	195	856
5. Planetary Impact (P5)	119	525	147	602

Analysis

We used the standardization (STD) method (Dorans & Holland, 1993) in conducting differential item, bundle, and passage analysis. The STD method is also referred to as the *p*-difference method in that it examines the difference in proportion correct on an item between test takers of different group membership but of matched ability. The STD method calculates the proportion correct on a given item between the focal and reference groups at each score level of the matching variable. Total test scores and scores on similar tests are common candidates for the matching variable. In this case, the TOEFL iBT reading total score was used as the matching variable for DIF analysis. The reading total score was selected because it best reflects TOEFL

iBT test takers' reading proficiency. In addition, no other standardized reading score was available for the large number of test takers included in this study.

The STD method provides a DIF index, referred to as the standardized *p*-difference (*STD P-DIF*). The following guidelines were used for interpreting DIF effect size in this study: $|STD\ P\text{-}DIF| < .05$ suggests a negligible DIF or A DIF; $.05 \leq |STD\ P\text{-}DIF| \leq .10$ suggests a moderate level of DIF or B DIF and may require further examination; and $|STD\ P\text{-}DIF| > .10$ suggests a substantial or C DIF and the item should be closely examined (Dorans & Holland, 1993; Penfield & Camilli, 2007).

Building on the same STD method, Dorans and Schmitt (1991, 1993) advanced the standardized mean difference (SMD) index for examining DIF for polytomously scored items. The ratio of SMD to its pooled standard deviation (SD_{SMD}) is used to indicate the DIF effect size of polytomous items: $|SMD/\ SD_{SMD}| < .17$ or not statistically significant from zero suggests A DIF; $.17 \leq |\ SMD/\ SD_{SMD}\ | \leq .25$ and significant from zero suggests B DIF; and $|\ SMD/\ SD_{SMD}\ | > .25$ and significant from zero suggests C DIF.

A purification procedure was applied to ensure that the matching variable was DIF-free (Zenisky, Hambleton, & Robin, 2003). After each DIF analysis, items identified with B or C DIF were removed from the calculation of the total reading score. The procedure was repeated until no DIF items contributed to the computation of the total reading score. Items deleted from each passage due to the purification procedure ranged from 1 to 4 out of a total of 14 items. The purified total reading score was used as the final matching variable for the DIF analysis.

In addition to investigating individual items, we are also interested in examining differential functioning for groups of items—item bundles. We constructed item bundles which consisted of items with either heavy presence of technical terms on the physical science passages or cultural information from outside North America on the cultural passages. If the concern that some TOEFL iBT reading items may favor test takers with certain disciplinary and cultural backgrounds is legitimate, then these item bundles are most likely to show differential functioning given the heavy presence of nonneutral information. The bundle approach has been shown to be a confirmative way to effectively detect differential functioning (Abbott, 2007; Douglas, Roussos, & Stout, 1996). The standardization method was also used for the DBF analysis. Again, the absolute values of DBF are organized into no or negligible (A), moderate (B), and large (C) DBF.

Finally, all the items on each testlet were treated as one polytomously scored item and were subject to a DIF analysis to see whether the testlet as a whole showed differential functioning. This analysis is referred to as the DPF. The statistical software PDIF (2006) developed by ETS was used to conduct analyses for DIF, DBF, and DPF.

When interpreting the results, we classified DIF, DBF, and DPF as uniform or nonuniform. Uniform differential functioning occurs when the direction and magnitude of DIF is the same across two proficiency levels, indicating that there is no interaction between test takers' content background knowledge and proficiency levels. Nonuniform differential functioning occurs when the focal (or reference) group is favored to a different degree across two proficiency levels, which is an indication of the interaction between content background knowledge and language proficiency. Three scenarios of non-uniform differential functioning were found in this study: (a) different directions of differential functioning across the two proficiency groups (e.g., an item favors the focal group for the higher proficiency group but favors the reference group for the lower proficiency group); (b) different magnitude of differential functioning across two proficiency levels (e.g., an item shows B level DIF for the higher proficiency group but C level DIF item for the lower proficiency group); and (c) presence of differential functioning for one proficiency group but absence of differential functioning for the other proficiency group.

Results

Item-Level Statistics

Table 4 presents the item difficulty index *p* and the item discrimination indices. The *p* of an item was calculated by averaging students' percentage scores on that item. Item discrimination was indicated by the biserial correlation between each item score and the TOEFL iBT reading test section scores, denoted as r_{bis}. Biserial correlation assumes normal distribution for the dichotomous variable, which is a more appropriate correlation index to use in this study. The statistics were obtained based on the information from all test takers who took the TOEFL iBT that contained the passages in our investigation.

The average item difficulty was similar across the five testlets relative to each of the test-taking groups. European Art (P1: $p = .63$), Planet Formation (P4: $p = .64$), and Planetary Impact (P5: $p = .62$) were relatively easy to their corresponding test-taking group and were very close in relative difficulty level. On average, students scored 60% or higher on the items based on these three passages. Restructuring of Japan (P2) was the most difficult testlet ($p = .49$) to its test

takers. It also had the most difficult item across the five testlets (p = .30). All the testlets had a relatively wide range of difficulty among their items.

Table 4

Item Difficulty and Item Discrimination Statistics for Items in Each Testlet

	p				r_{bis}			
	M	*SD*	Min	Max	*M*	*SD*	Min	Max
1. European Art (P1)	.63	.11	.46	.83	.56	.08	.44	.66
2. Restructuring of Japan (P2)	.49	.13	.30	.69	.47	.10	.37	.69
3. Painting in Japan (P3)	.59	.17	.37	.84	.48	.08	.30	.60
4. Planet Formation (P4)	.64	.15	.33	.88	.50	.07	.39	.60
5. Planetary Impact (P5)	.62	.12	.38	.75	.49	.11	.23	.70

Item discrimination indices were satisfactory in general. The average biserial correlations of the items on each testlet with the TOEFL iBT reading scores were reasonably high, ranging from .47 to .56. There was only one item that had an r_{bis} below .25, and it belonged to the Planetary Impact (P5) testlet. This is a relatively difficult item, and only 38% of test takers answered it correctly.

Differential Item Functioning (DIF)

Table 5 shows the results for the DIF analysis for all five testlets. Each item was analyzed twice for DIF analysis: once for the lower proficiency level test takers and once for the higher. In total, 140 DIF analyses were conducted (14 items on each passage x five passages x two proficiency groups).

The proportions of B- and C- level DIF items were similar between the lower (17.14%, 12 items) and higher proficiency test takers (18.57%, 13 items) among all 70 items (Table 5). The number of items that showed DIF for either group was 17 (24.28%), among which only seven items showed the same level of DIF in the same directions for both proficiency groups. These seven items were uniform DIF items and did not suggest interactions between background knowledge and proficiency. The rest of the 10 were then nonuniform DIF items, which indicates an interaction between background knowledge and proficiency level.

Table 5

Number of Different Levels of Differential Item Functioning (DIF) Items for Lower and Higher Proficiency Students

	Lower proficiency (*N*)					Higher proficiency (*N*)					DIF items	Overlapped
	A	B		C		A	B		C		for either	DIF items for
Passages		+	-	+	-		+	-	+	-	group	both groups
1. European Art (P1)	12	1	1	0	0	10	1	1	1	1	5	1
2. Restructuring of Japan (P2)	10	2	2	0	0	11	1	2	0	0	4	3
3. Painting in Japan (P3)	11	1	2	0	0	10	2	2	0	0	4	3
4. Planet Formation (P4)	13	0	1	0	0	13	1	0	0	0	2	1
5. Planetary Impact (P5)	12	1	1	0	0	13	1	0	0	0	2	1

Among the 17 DIF items, only two were of C-level and the rest were B-level DIF items for either proficiency group. These DIF items did not consistently favor the focal group. Both C-level DIF items were identified for the higher proficiency group as on the European Art (P1) testlet: one favoring the focal group and one favoring the reference group. There were 12 B-level DIF items for the lower proficiency group and 11 for the higher proficiency group across the five passages. For the lower proficiency group, six B-level DIF items favored the reference group and six items favored the focal group. For the higher proficiency group, five B-level DIF items favored the reference group and six favored the focal group.

DIF items were presented in Figures 1 and 2. A bar pointing to the right favored the focal group and one to the left the reference group. A level DIF items were not present in the figures. The lower proficiency group was represented by patterned bar and the higher proficiency group the black bar.

Figure 1 showed that for three out of the five testlets (European Arts [P1], Planet Formation [P4], and Planetary Impact [P5]), DIF items were of different magnitudes between the two proficiency groups, though the direction of the DIF was the same for both groups, which suggests that whether an item favored the reference or the focal group was consistent across the two proficiency groups. The DIF patterns were the most consistent between the two proficiency groups on Restructuring of Japan (P2) and Painting in Japan (P3). In each testlet, three of four DIF items were of the same direction and magnitude between the two proficiency groups. It was interesting to notice that while the passage on the European culture (European Art [P1] testlet) generated the most DIF items and largest interaction between DIF occurrence and proficiency

level, the two passages related to Japanese culture had the greatest proportion of DIF items that overlapped between two proficiency groups (75%), thus showing minimal interaction between background and proficiency. The two science passages had relatively few DIF items (two in each testlet) in comparison to the cultural passages.

The results suggest that the interaction between differential functioning and language proficiency depended on the content of the passage and items.

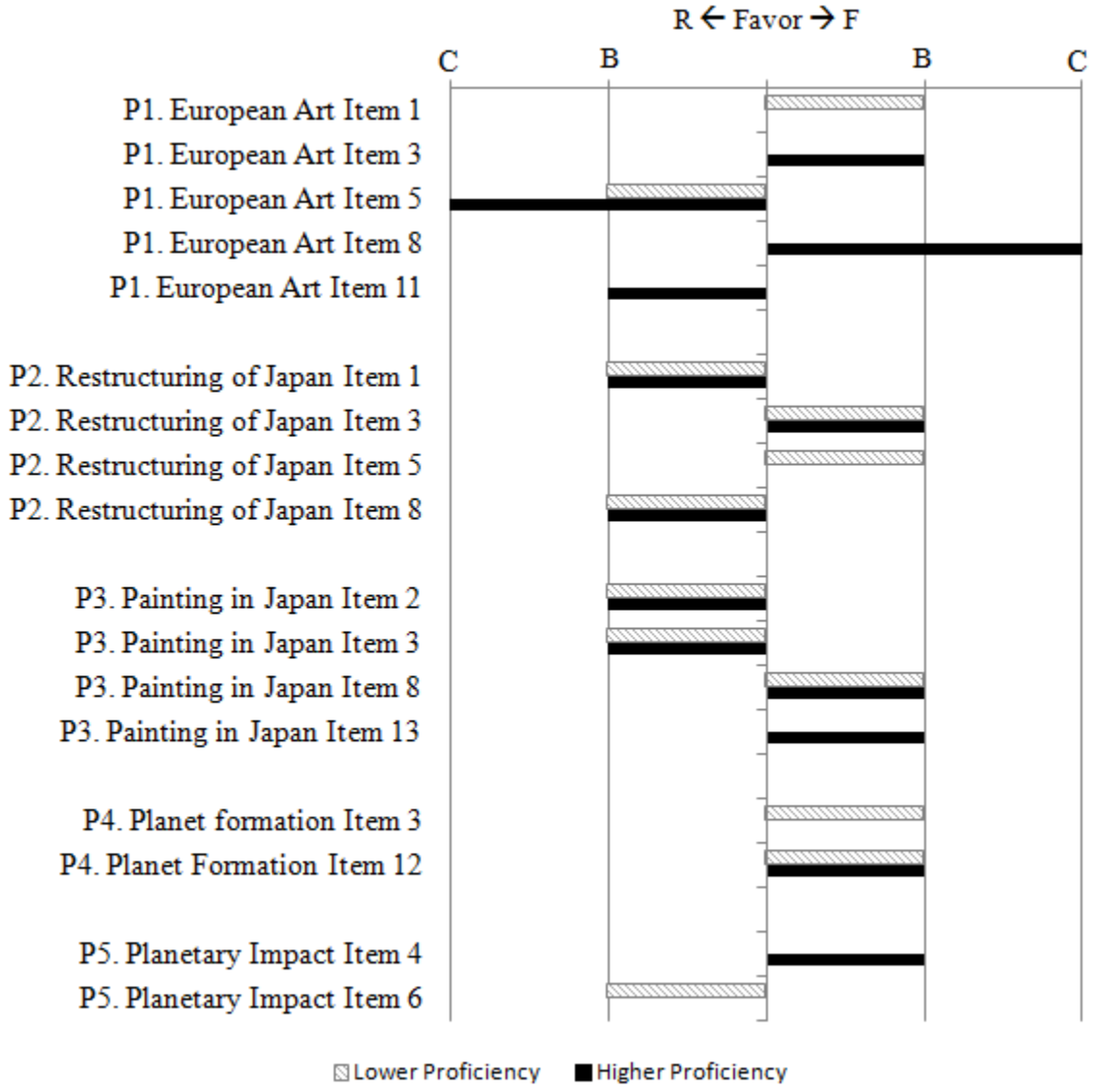

Figure 1. **Classification of differential item functioning (DIF) items.**

Note. B = B-level DIF items, C = C-level DIF items, F = focal group, R = reference group.

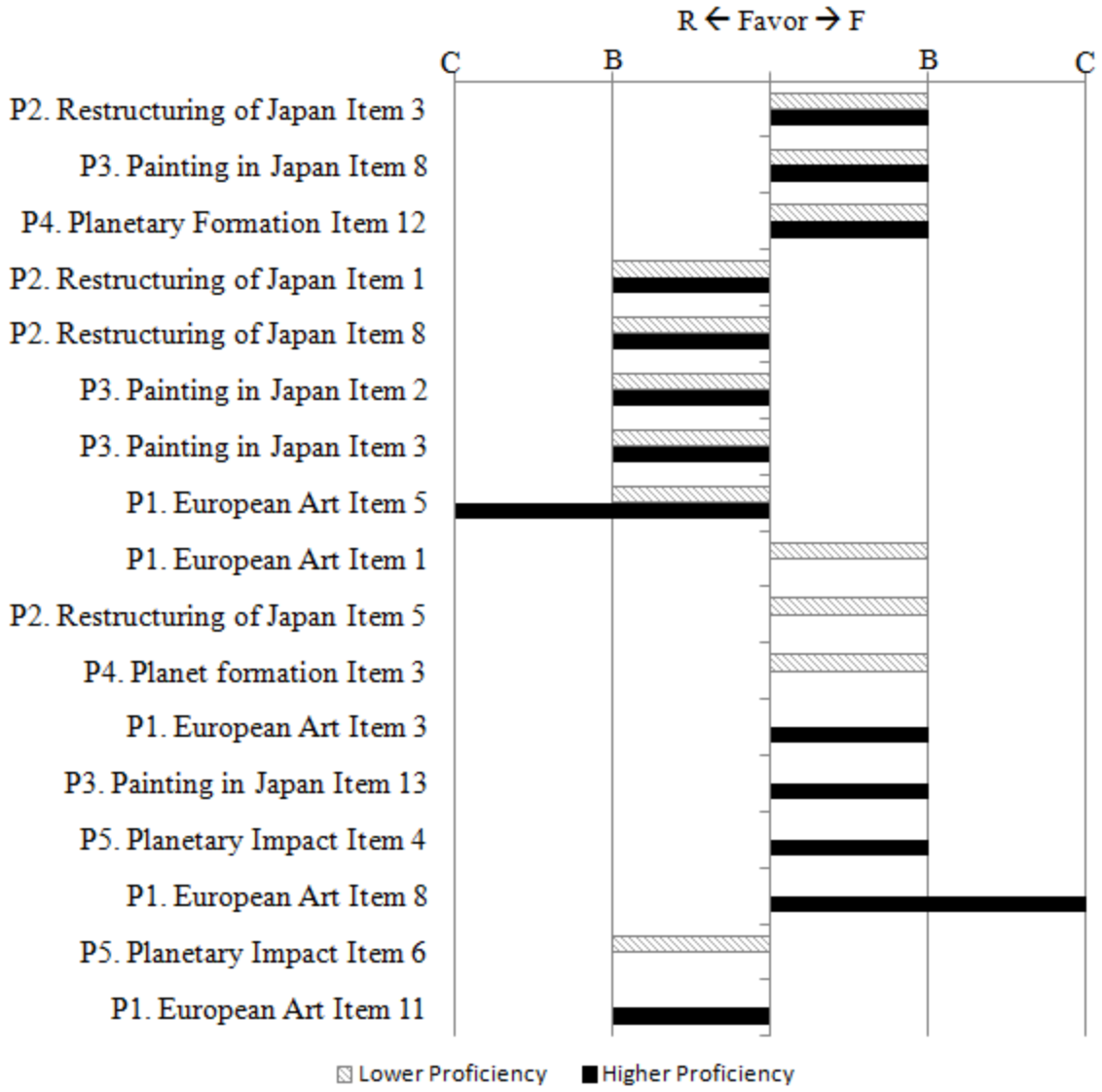

Figure 2. **Classification of DIF items by uniformity.**

Note. B = B-level DIF items, C = C-level DIF items, F = focal group, R = reference group.

In Figure 2, the DIF items are grouped based on their uniformity across the two proficiency groups. The DIF items that are of the same direction and magnitude are presented at the top, followed by the rest. From Figure 2, we identified three patterns of DIF occurrence between the two proficiency groups: (a) uniform DIF items of the same magnitude and same direction between the two proficiency groups, (b) nonuniform DIF items of the same direction

but of different magnitudes, and (c) nonuniform DIF items that exist for only one of the proficiency groups. We did not observe any opposite direction DIF items across the two proficiency groups. In the study, 7 of the 17 DIF items were uniform DIF items, indicating that these items favored or disfavored higher proficiency and lower proficiency students with background knowledge to the same degree. Item 5 on the passage European Art (P1) exhibited C-level DIF for the higher proficiency group and B-level DIF for the lower proficiency group, both in favor of the reference group. The finding means that this item favored both the lower and higher proficiency students who did not have the European culture background knowledge, and the item favored the higher proficiency students to a greater degree. Nine items showed DIF for only one of the proficiency groups, suggesting that background knowledge only manifested its effect on one proficiency group.

Differential Bundle Functioning (DBF) and Differential Passage Functioning (DPF)

DBF analysis was also conducted separately for the lower and higher proficiency groups (see Table 6). Note that the individual items identified with DIF in a passage may or may not have been included for the DBF analysis depending on whether the items met the selection criterion for the bundle construction. Similarly, items included in the bundle may not show DIF as individual items. The first column in Table 6 is the total number of items in each of the five passages, followed by the number of items in each bundle. The last two columns show the level of DBF for each proficiency group. Only the Planet Formation (P4) had a B-level DBF for both proficiency groups, favoring the focal group in each case. Other testlets exhibit no DBF. Since the direction and the magnitude of the DBF were the same for both proficiency groups, the interaction effect between background knowledge and language proficiency was not evident at the bundle level for the testlets investigated.

No differential passage functioning (DPF) was found when all the items on each passage were analyzed as one polytomously scored item, which suggests that background knowledge does not have an impact on student performance at the reading passage level.

Table 6

Information on Item Bundles and Differential Bundle Functioning (DBF) for Each Testlet

Passages	Total number of items	Number of items in the bundle	DBF	
			L.P.	H.P.
1. European Art (P1)	14	7	--	--
2. Restructuring of Japan (P2)	14	4	--	--
3. Painting in Japan (P3)	14	8	--	--
4. Planet Formation (P4)	14	6	+ B	+ B
5. Planetary Impact (P5)	14	8	--	--

Note. +B = B-level DBF, H.P. = higher proficiency group, L.P. = lower proficiency group "—" represents the absence of DBF.

Discussion

We set forth to investigate whether items on the TOEFL iBT reading test favored test takers of certain major fields of study and with certain cultural backgrounds, with the intention of factoring in the interaction between disciplinary and cultural effects and the test takers' language proficiency level.

The items investigated were of moderate difficulty (p = .49 to .64) and satisfactory item discrimination (r_{bis} = .25 - .56).

Differential Item Functioning (DIF) of Different Directions Observed for Both Proficiency Levels

Among the 70 items investigated across the five testlets, 17 DIF items occurred in either the higher or lower proficiency group. Among them, only seven (41.18%) items had DIF of the same direction and magnitude across the two proficiency groups. Ten items favored the focal group for either proficiency group and seven items favored the reference group for either proficiency group. A close examination of the data in Table 5 and Figures 1 and 2 revealed that in the lower proficiency group six items favored the focal group, and another six items favored the reference group. In the higher proficiency group, seven items favored the focal group, and six items favored the reference group.

This finding provides evidence that background knowledge does not always work to one's advantage, regardless of one's proficiency level; though in this case, more items favored the focal group than the reference group. A plausible explanation for why some of the DIF items

disadvantaged the focal group could be that the test takers in the focal group relied too heavily on their background knowledge in answering multiple-choice reading comprehension items. They might have chosen an option which, while culturally appropriate or correct, did not display the relationship or logic that the item asked for. On the TOEFL reading test, understanding the local message (clausal, sentential level parsing) and using vocabulary and syntactic linguistic decoding were as important as understanding the global message (within-paragraph, across-paragraph processing) and applying content background knowledge to making a meaningful interpretation of the textual information, if not more so. For example, Jang (2009) reported that among 37 reading comprehension questions in a prototype of TOEFL iBT in the *LanguEdge*™ courseware, through think-aloud protocol with 11 ESL students in TOEFL preparation courses, 17 items were identified as engaging the reading processes of inferencing, summarizing, and mapping contrasting ideas into mental framework, which supposedly entailed top-down heavy processing. Nevertheless, 12 out of these 17 items were also found to engage processes of vocabulary, syntactic and semantic linking (including understanding negation), and understanding textually explicit information reading processes, which supposedly entailed bottom-up driven processing. This indicates that bottom-up processing is very important in TOEFL iBT reading comprehension, and if a reader exercises top-down or content-knowledge-driven processing to the extent of overlooking the bottom-up or text-driven processing, performance might be disadvantaged. Carrell (1988) also noted that second language readers, compared to native-speaker readers, tended to rely too heavily on culturally biased schemata, which interfered with their text-based comprehension.

The explanation for why the focal group was disadvantaged on reading items has several implications for test developers. First, when including culturally or content-correct information in distracters, test developers need to consider the potential impact that such distracters may have on test takers who possess such cultural or content knowledge. One way to investigate this is through differential distracter functioning (DDF) analysis, to determine if different groups of test takers with equal ability are differentially attracted to the distracters on a multiple-choice item that present correct cultural or subject matter information but are wrong answers to the question (Banks, 2006).

Second, test developers need to distinguish whether DIF found is construct-relevant (the appropriate part of the intended construct) or construct-irrelevant (an unintended part of the

construct dimension). If the DIF item is construct-relevant, test takers who score higher on the item do so not because they used construct irrelevant strategies, such as guessing, or test-wise strategies, but because their general linguistic knowledge and skills were superior. We might use think-aloud protocols to examine test takers' comprehension processing and test-taking processes in the focal and reference groups. For example, Cohen and Upton (2006), using think-aloud protocols, found that in taking the practice TOEFL iBT, students at the advanced proficiency level made educated guesses using background knowledge or extratextual knowledge. Pritchard (1990) found that 11th grade readers used the strategies of establishing intrasentential ties and using background knowledge more often when processing culturally familiar passages than when processing culturally unfamiliar ones. By examining think-aloud data, Jang (2009) found that the activation of content background knowledge was associated with items that required inferencing skills. Even when ESL learners engaged inferencing skills and connected textual information to their background knowledge, the utilization of background knowledge did not always lead to successful performance. Jang discovered that inferencing skills were "more cognitively demanding and required test takers to use multiple skills simultaneously" (p. 225), which requires learners to connect lexical, syntactic, and rhetorical knowledge, in addition to applying content background knowledge to achieve global level understanding. Similar investigations should be conducted for the TOEFL iBT test-taking population, with further analysis of the conditions under which background knowledge related strategies become advantageous or disadvantageous. The investigations should answer questions such as under what combination of text and task conditions a student with or without the assumed background knowledge would perform best on a reading test item. Test features include syntactic complexity, text cultural information load, text length, and so on. Task characteristics can be levels of inferencing skills required by the item, item difficulty, distracter difficulty, and distracter alignment with cultural background.

Consistent Differential Item Functioning (DIF) Direction but Varying DIF Magnitude Between the Lower and Higher Proficiency Groups

The direction of the DIF items was very similar for both proficiency groups, which means if an item favored either the focal or the reference group for test takers from the lower proficiency level, it tended to favor the same group for those from the higher proficiency level. No items favored opposite groups (focal vs. reference) across the lower and higher proficiency

groups. This showed that background knowledge either advantaged students from both proficiency levels or disadvantaged both proficiency groups. We did not observe cases where items showed DIF with opposite directions across the higher and lower proficiency groups.

Although consistent in the direction of DIF patterns, the two proficiency groups showed differences in the magnitude of DIF. Nine items only showed DIF for one group, not the other. One item (P1 Item 5) showed varying magnitude between the two proficiency groups. In the European Art (P1) testlet in particular, the magnitude of DIF differed between two proficiency groups on every DIF item.

Although the difference in DIF magnitude between the two proficiency groups was at most one level apart (A-B, or B-C), there was one item (Item 5) on the European Art (P1) testlet that showed a two-level difference: It showed large (C-level) DIF for the higher proficiency group but low (A-level) DIF for the lower proficiency group.

Two items (Item 5 and Item 11) on European Art (P1) are noteworthy. Both items favored the reference group, and the DIF magnitude was larger for the higher proficiency focal group than for the lower proficiency focal group. These two items were the easiest items in this testlet, but with reasonable discrimination power. The *p* values for Items 5 and 11 were .83 and .77, respectively, and the r_{bis} correlations were .64 and .48, respectively. Easy items like these may require more text-based processing and less knowledge-based processing. The fact that the focal group with higher proficiency was disadvantaged to a greater degree than the lower proficiency focal group could be because the individuals in the higher proficiency focal group over-exercised knowledge-based processing compared to their reference group counterparts and therefore were probably more susceptible to some distracters that were culturally relevant but were not related to what the items targeted.

In summation, we can see that passage- and item-specific characteristics are both at work in the interaction between background knowledge and proficiency level. The physical science passages had fewer DIF items than the passages with cultural content. The reason may be that the content in those two passages—Planet Formation (P4) and Planetary Impact (P5)—was more general than the content in the culture passages. Clapham (1996) also identified an interaction between background knowledge and passage types on the IELTS test, which she attributed to the effect of the content specificity of the passage. If the content is more general, it requires less background knowledge, and therefore the effect of background knowledge would not be expected

to be as evident as when the content is specific. On the contrary, if the passage, or the content tested by an item, is specific, then the background knowledge is expected to be more active in the reading process for the focal group, and therefore, the difference between the focal and the reference group would be expected to be larger. In language testing, the specificity of the passage and that of the content tested by individual items may both play a role in contributing to the effect of background knowledge. Other test task variables suggested by Alderson (2000) may also be at work—variables that are possibly related to item format, item stem content, or item sequence.

Differential Bundle Functioning (DBF) and Differential Passage Functioning (DPF)

Five item bundles were constructed, composed of the items deemed to have heavy presence of cultural or domain-specific terminology. Out of the five item bundles, only the Planet Formation (P4) bundle showed B-level differential functioning favoring the focal group. This result is interesting because the Planet Formation (P4) testlet only had two DIF items at the individual item level, and only one of them was in the bundle. This suggests that although individual items in a bundle may not show differential functioning, they, as a group, may have cumulatively functioned differentially in favor of the focal or reference group. DBF analysis provided additional insight into an aggregated effect of background knowledge.

No DPF was found for any of the testlets. The result of the DBF and DPF was probably due to the fact that some items favored the focal group and some the reference group; for most testlets, the effect of background knowledge was cancelled out. Therefore, except for Planet Formation (P4), in which all the DIF items favored the focal group, none of the other testlets showed differential bundle functioning or differential passage functioning.

Next Steps in Research

We propose several future research topics to further our understanding of the effect of background knowledge in language testing and its impact on test fairness. First, we might use alternative approaches to investigate the existence of nonuniform DIF, such as logistic regression and item response theory (IRT) modeling. Second, think-aloud research may provide in-depth information about the underlying reasons for certain students' choices in their item responses on the reading test. Third, future research may benefit from content analysis using tools such as components of communicative language ability and test method facet taxonomic framework (Bachman, Davidson, Ryan, & Choi, 1995). In addition, DDF, as described by Banks (2006),

would also contribute to the understanding of DIF existence.

Conclusion

This study investigated the interaction effect of background knowledge and language proficiency on TOEFL iBT reading passages. Three cultural passages and two physical science passages and their items were selected for investigation. The majority of the items showed no DIF. Some evidence of interaction between background knowledge and proficiency level at the item level exists, as 10 out of 17 DIF items showed different magnitudes in differential functioning between the two proficiency groups. The interaction effect was more evident in some passages than in others, pointing to the fact that idiosyncratic characteristics of both passages and items play a role in DIF existence, such as specificity of the passages and the content tested by certain items. Among all the testlets, interaction effect was most evident in European Art (P1), which had 5 out of 10 non-uniform DIF items and also had DIF items of the largest magnitude. When items from this passage were excluded, the interaction effect between background knowledge and language proficiency was not very evident—less than 10% (5 out of 56) of the remaining items showed nonuniform DIF.

No evidence of interaction between background and proficiency level was observed for the bundle-level analysis. The Planet Formation (P4) passage was the only passage that showed B-level DBF in favor of the focal group for both proficiency groups.

When examined holistically, the TOEFL iBT reading passages were neither advantageous nor disadvantageous to those who had physical science backgrounds or were familiar with a certain culture, and this holds for both the lower and higher proficiency groups. This finding adds evidence to the argument in support of the validity of the TOEFL iBT, in that the testlet scores on theses passages containing subject content and cultural information were not biased against any knowledge group. The results provide evidence that it is possible for the TOEFL iBT to create reading passages that are contextualized in subject and cultural material but at the same time free from bias at the passage level.

The findings in this study also advance our understanding of the effect of background knowledge in reading comprehension, specifically, the understanding gained in Liu et al.'s (2009) study. Both studies found no passage-level differential functioning. Through careful and rigorous content and qualitative bias control, test developers can create a reading comprehension task free of measurable background knowledge bias on the overall task level. The current study

also provided more fine-grained results than those found in Liu et al.. For example, Liu et al. identified two B-level and one C-level DIF items in the passage on European Art (P1), both favoring the focal group. However, in the current study, four DIF items were identified for the higher proficiency group and two DIF items were found for the lower proficiency group, and only half of the items favored the focal group. Therefore, consideration of the interaction between language proficiency and background knowledge presents an in-depth picture of the effect of background knowledge on test performance, which promises to be informative for test developers as they examine TOEFL iBT reading comprehension items from the point of view of test fairness.

This research also contributes to the theory of the effect of background knowledge in second language reading comprehension. It calls to researchers' attention that when background knowledge effect manifests itself on comprehension tasks, its effect, both in magnitude and in direction, may vary across tasks, items, and different proficiency groups. Therefore, it is of central importance to thoroughly investigate the conditions when background knowledge effects test performance.

References

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

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Alderson, J. C., & Urquhart, A. H. (1985a). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, *2*, 192–204.

Alderson, J. C., & Urquhart, A. H. (1985b). This test is unfair: I'm not an economist. In P. C. Hauptman, R. Leblanc, & M. B. Wesche (Eds.), *Second-language performance testing* (pp. 25–43). Ottawa, Canada: University of Ottawa Press.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I-C. (1995). *The Cambridge-TOEFL comparability study: Final report.* Cambridge, UK: University of Cambridge Local Examinations Syndicate.

Banks, K. (2006). A comprehensive framework for evaluating hypotheses about cultural bias in educational testing. *Applied Measurement in Education*, *19*, 115–132.

Breland, H., Lee, Y-W., Najarian, M., & Muraki, E. (2004). *An analysis of the TOEFL CBT writing prompt difficulty and comparability of different gender groups* (TOEFL Research Report No. TOEFL-RR-76). Princeton, NJ: ETS.

Brown, J. D. (1984). A norm-referenced engineering reading test. In A. K. Pugh & J. M. Ulijn (Eds.), *Reading for professional purposes* (pp. 213–222). London, UK: Heinemann Educational Books.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Carrell, P. L. (1988). SLA and classroom instruction: Reading. *Annual Review of Applied Linguistics*, *9*, 223–242.

Chan, C. Y. H. (2003). Cultural content and reading proficiency: A comparison of mainland Chinese and Hong Kong learners of English. *Language, Culture and Curriculum, 16*, 60–69.

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

Chen, Q., & Donin, J. (1997). Discourse processing of first and second language biology texts: Effects of language proficiency and domain-specific knowledge. *The Modern Language*

Journal, *81*, 209–27.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge, UK: Cambridge University Press.

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph Series No. TOEFL-MS-33). Princeton, New Jersey: ETS.

Davidson, B. (2004). *Comparability of test scores for non-aboriginal and aboriginal students* Unpublished doctoral dissertation). University of British Columbia: Ottawa, Canada.

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 Associates.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report 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* (p. 135–165). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Elder, C., McNamara, T., & Congdon, P. (2003). Rasch techniques for detecting bias in performance tests: An example comparing the performance of native and non-native speakers on a test of academic English. *Journal of Applied Measurement, 4,* 181–197.

Erickson, M., & Molloy, J. (1983). ESP test development for engineering students. In J. W. 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*, 113–148.

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement*, *64*, 925–936.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.

Goodman, K. S. (1976). Reading: A psycholinguistic guessing game. In H. Singer & R. B. Rudell (Eds.), *Theoretical models and processes of reading* (pp. 497–508). Newark, DE: International Reading Association.

Hale, G. A. (1988). Student major field and text content: Interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing*, *5*, 49–61.

Hammadou, J. (1991). Interrelationships among prior knowledge, inference, and language proficiency in foreign language reading. *The Modern Language Journal*, *75*, 27–38.

Hock, T. S. (1990). The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates. In J. H. A. L. D. Jong and D. K. Stevenson (Eds.), *Individualizing the assessment of language abilities* (pp. 214–224). Clevedon, PA: Multilingual Matters.

Hudson, T. (2007). *Teaching second language reading*. Oxford, UK: Oxford University Press.

Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, *6*, 210–238.

Jang, E. E., & Roussos, L. (2009). Integrative analytic approach to detecting and interpreting L2 vocabulary DIF. *International Journal of Testing*, *9*, 238–259.

Johnson, P. (1981). Effects on reading comprehension of language complexity and cultural background of a text. *TESOL Quarterly*, *15*, 169–181.

Keshavarz, M. H., Mahmoud, R. A., Ahmadi, H. (2007). Content schemata, linguistic simplification, and EFL reader's comprehension and recall. *Reading in a Foreign Language*, *19*, 19–33.

Kim, Y., & Jang, E. E. (2009). Differential functioning of reading subskills on the OSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. *Language Learning*, *59*, 825–865.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163–182.

Koh, M. Y. (1985). The role of prior knowledge on reading comprehension. *Reading in a Foreign Language*, *3*, 375–380.

Lee, Y-W., Breland, H., & Muraki, E. (2005). Comparability of TOEFL CBT writing prompts for different native language groups. *International Journal of Testing*, *5*, 131–158.

Liu, O. L., Schedl, M., & Kong, N. (2009). *Does content knowledge affect TOEFL iBT reading*

performance? A confirmatory approach to differential item functioning (TOEFL iBT Research Series Report No.TOEFLiBT-09). Princeton, NJ: ETS.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, England: Basil Blackwell.

Pae, T. (2004). DIF for learners with different academic backgrounds. *Language Testing*, *21*, 53–73.

Pae, T., & Park, G. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing*, *23*, 475–496.

Park, G. (2004). Comparison of L2 listening and reading comprehension by university students learning English in Korea. *Foreign Language Annals*, *37*, 448–458.

PDIF: Computer 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, NY: Elsevier.

Peretz, A. S., & Shoham, M. (1990). Testing reading comprehension in LSP: Does topic familiarity affect assessed difficulty and actual performance? *Reading in a Foreign Language*, *7*(1), 447–454.

Pritchard, R. (1990). The effects of cultural schemata on reading processing strategies. *Reading Research Quarterly*, *25*, 273–295.

Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly*, *4*, 165–189.

Roussos, L., & Stout, W. (1996). A multidimensionality-biased DIF analysis paradigm. *Applied Psychological Measurement*, *20*, 273–295.

Roznowski, M., & Reith, J. (1999). Examining the measuring quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement*, *59*, 248–269.

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

Salager-Meyeer, F. (1994). Reading medical English abstracts: A genre study of the interaction between structural variables and the reader's linguistico-conceptual competence. *Journal*

of Research in Reading, *17*, 120–146.

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

Singh, M. D. (2005). *Investigating the prediction of item difficulty and bias by gender in reading literacy* (Unpublished doctoral dissertation). Ontario Institute for Studies in Education of the University of Toronto: Ottawa, Canada.

Smith, F. (1994). *Understanding reading: A psycholinguistic analysis of reading and learning to read*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Snetzler, S., & Qualls, A. L. (2000). Examination of differential item functioning on a standardized achievement battery with limited English proficiency students. *Educational and Psychological Measurement*, *60*, 564–577.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, *17*, 323–340.

Usó-Juan, E. (2006). The compensatory nature of discipline-related knowledge and English language proficiency in reading English for academic purposes. *The Modern Language Journal*, *90*, 210–227.

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*, 51–64.

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