ETS TOEFL

Quality Beyond Measure.

TOEFL Junior® Research Report
TOEFL JR–01

Assessing the Test Information Function and Differential Item Functioning for the *TOEFL Junior*® Standard Test

John W. Young

Rick Morgan

Paul Rybinski

Jonathan Steinberg

Yuan Wang

September 2013

Assessing the Test Information Function and Differential Item Functioning for the *TOEFL Junior*® Standard Test

John W. Young, Rick Morgan, Paul Rybinski, Jonathan Steinberg, and Yuan Wang
Educational Testing Service, Princeton, New Jersey

RR-13-17

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 Junior*® Standard Test is an assessment that measures the degree to which middle school-aged students learning English as a second language have attained proficiency in the academic and social English skills representative of English-medium instructional environments. The assessment measures skills in three areas: listening comprehension, language form and meaning, and reading comprehension. This study focused on two specific psychometric characteristics of the assessment: (a) For which segments of its score scales does the assessment provide sufficient information to support the use of the scores in placement decisions? (b) Do items exhibit significant (or C-level) differential item functioning (DIF) when comparisons are made between test-takers from different countries? For the first question, both of the forms we analyzed appear to provide sufficient information to support placement decisions across the majority of the score scale for all 3 sections of the assessment. For the second question, we found that a moderate number of items exhibited significant DIF, while the linguistic analyses conducted on the DIF results showed plausible construct-relevant explanations for most of the findings.

Key words: TOEFL Junior, language proficiency, language assessment, English proficiency test, middle school students

The *TOEFL*® test 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, the *TOEFL iBT*® test. 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. The *TOEFL Junior*® tests, which measure the English proficiencies of English language learners aged 11 to 15, was released in 2012. With it emerges a need for a separate, but related research series. Beginning in 2013, a TOEFL Junior research report series was launched.

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

Current (2012-2013) members of the TOEFL COE and the TOEFL Young Learners Subcommittee are:

TOEFL COE

John M. Norris – Chair, Georgetown University
Maureen Burke, The University of Iowa
Yuko Goto Butler, University of Pennsylvania
Barbara Hoekje, Drexel University
Ari Huhta, University of Jyväskylä, Finland
Eunice Eunhee Jang, University of Toronto, Canada
James Purpura, Teachers College, Columbia University
John Read, The University of Auckland, New Zealand
Carsten Roever, The University of Melbourne, Australia
Steve Ross, University of Maryland
Norbert Schmitt, University of Nottingham, UK
Ling Shi, University of British Columbia, Canada

TOEFL Young Learners Subcommittee

Frances Butler, Language Testing Consultant
Yuko Goto Butler, University of Pennsylvania
Anna Chamot, The George Washington University
Barbara Hoekje, Drexel University
Eunice Eunhee Jang, University of Toronto, Canada
Lorena Llosa, New York University
Steven Ross, University of Maryland

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

The *TOEFL Junior*® Standard Test is an assessment that measures the degree to which middle school students have attained proficiency in the academic and social English skills representative of English-medium instructional environments. It is a general English proficiency assessment that is not tied to any specific curriculum. The assessment measures skills in three areas: listening comprehension, language form and meaning, and reading comprehension. Each skill area is assessed within its own separate section of the TOEFL Junior Standard test.

Within the Listening Comprehension section of the TOEFL Junior Standard test, the following three types of listening abilities are assessed:

- The ability to listen for basic, interpersonal purposes: Students must be able to comprehend conversations about day-to-day matters on familiar topics that take place in school settings. This includes the ability to understand main ideas and important details, the ability to make inferences based on what is implied but not explicitly stated and to make predictions, the ability to understand the speaker's purpose, and the ability to correctly interpret such features of spoken language as intonation and contrastive stress.
- The ability to listen for instructional purposes: Students must be able to comprehend the language that teachers and other school staff make for a range of purposes *other* than presenting academic content. This includes language that takes place both inside and outside of the classroom (e.g., on field trips or in the school library or auditorium) and that fills a range of speech functions (e.g., making announcements, giving reminders, issuing invitations, making warnings). In listening to such oral language, students must be able to understand the main idea of a message, to identify a speaker's purpose, to make inferences based on what is implied but not explicitly stated, and to make predictions based on what the speaker says.
- The ability to listen for academic purposes: Students need to comprehend ideas presented in a lecture or discussion that is based on academic material. These lectures or discussions will range in level of formality and will reflect the features typical of oral language (e.g., relatively complex verb structures, relatively little nominalization, and occasional performance disfluencies).

The Language Form and Meaning section of the TOEFL Junior Standard test assesses the enabling skills of grammar and vocabulary. While communicative skills such as listening and

reading are of primary importance in developing the English proficiency needed for English-medium instructional environments, enabling skills such as grammar and vocabulary also play a significant role for students learning in international contexts. These enabling skills are not ends in and of themselves, but these are important in contributing to and helping to develop communicative skills. In developing enabling skills, it is important that students not focus exclusively on form but rather learn the forms of English as a tool to create a range of meanings in a variety of contexts. Students need to be able to recognize grammatical forms and lexical items that are correct and that create appropriate meanings in a range of syntactic and cohesive contexts (e.g., based both on sentence grammar and on broader discourse). They need to be able to recognize these forms and meanings in a range of school-based texts (e.g., informational texts, brochures and advertisements, student writing).

Within the Reading Comprehension section of the TOEFL Junior Standard test, the following two types of reading skills are assessed:

- The ability to read and comprehend academic texts: Students need to be able to read and comprehend academic texts in a range of genres (e.g., expository, biographical, persuasive, literary) across a range of subject areas (e.g., arts and humanities, science, social science). They need to be able to read such texts at difficulty levels up to and including those typical of what is used in English-medium classrooms. In reading these texts, students need to be able to understand main ideas and key supporting information, to make inferences based on what is implied but not explicitly stated, and to understand key vocabulary (either from previous knowledge or from context) and cohesion within the text (e.g., referential relationships across sentences). Depending on the nature of the specific text, students may also need to understand the author's purpose, follow the logic and intended meaning of basic rhetorical structures, follow steps or directions, and/or identify and understand figurative language. As with listening, reading texts should not require any specific background knowledge, but will sometimes require students to read in order to learn new information in an academic context.
- The ability to read and comprehend nonacademic texts: While academic texts pose the primary reading challenge in English-medium instructional environments, students must also be able to read a variety of types of nonacademic texts. These

include correspondence (e.g., e-mail messages, letters), journalism, and student writing, as well as forms of text that tend to be less linear (e.g., brochures, menus, advertisements, schedules). In reading nonacademic texts, students not only must demonstrate the same types of understanding detailed above for academic texts, but also that they understand features that are more typical of nonacademic texts (e.g., greater use of idiomatic language).

The TOEFL Junior Standard test is an assessment that requires 110 minutes of testing time. The assessment contains a total of 126 multiple-choice items divided equally among the three sections. Each section score is reported on a scale from 200 to 300; the total score is a sum of the three section scores and thus ranges from 600 to 900. The first operational administration of the TOEFL Junior Standard test was given in October 2010.

This research study focused on two specific psychometric characteristics of the TOEFL Junior Standard test:

1. For which segments of its score scales does the assessment provide sufficient information to support the use of the scores in placement decisions?
2. Do items exhibit significant differential item functioning (DIF) when comparisons are made between test-takers from different countries?

In this study we are investigating both the test information functions and DIF for the TOEFL Junior Standard test because these two research questions provide distinct, but related information, on the technical qualities of the assessment. Analysis of the test information functions provides evidence on the overall performance of the assessment across the score scale for each section. The DIF analyses provide evidence on the quality of the assessment at the item level; items that do not function effectively will have a detrimental impact on the test information functions for the assessment.

Relevant Prior Research

Differential Item Functioning in International English Language Proficiency Tests

DIF studies have been conducted on many large-scale international English language proficiency tests. Among them, the most frequently used grouping variables are gender (Aryadoust, Goh, & Kim, 2011; Breland, Lee, Najarian, & Muraki, 2004) and native language background. Because gender, and some less commonly used variables such as academic major and familiarity

with a particular culture (Liu, Schedl, Malloy, & Kong, 2011), are not of interest in this study, we focused on the studies that evaluated item functioning across native language groups.

One of the earliest DIF studies on TOEFL (Alderman & Holland, 1981) investigated item performance across six native language groups: African, Arabic, Chinese, Germanic, Japanese, and Spanish, with subgroups for some of them. Significant differences were found on over 80% of the items. Based on the results of a content analysis conducted by English-as-a-second-language (ESL)/English-as-a-foreign-language expert reviewers, researchers suggested that performance on given items in a test of proficiency in a second language would vary according to linguistic contrasts with test-takers' native languages. In other words, the relative advantage of a language group on a specific item could be attributed to the linguistic similarity of that language to English. However, when inspecting a test form and the answer key alone, reviewers could not identify which items would exhibit differential performance across language groups. Hence, the researchers called for statistical procedures necessary for identifying items with exaggerated or unexplained differences.

Similar implications were suggested in Ryan and Bachman's (1992) quantitative analysis of both the TOEFL test and First Certificate of English (an upper-intermediate level English proficiency test developed by Cambridge ESOL), adopting a more general approach of dividing participants into Indo-European (IE) and Non-Indo-European (NIE) language backgrounds. The TOEFL Listening, Vocabulary and Reading Comprehension, and Structure and Written Expression sections were found to include DIF items favoring either group. The First Certificate of English Listening section included some DIF items that favored the IE group. Without a comprehensive content analysis, the researchers suggested that spoken language might be more sensitive to similarities between test-takers' native languages and the target language in testing. The reasons that some TOEFL items favored the NIE group were not accounted for, although one important factor could be the curriculum focus of different countries (e.g., East Asian countries focus more on the teaching of grammar, so there might be some Structure and Written Expressions items that favored the NIE group). For another well-known international English test, the International English Language Testing System, Aryadoust (2012) administered a practice form of the Listening section to 209 participants from Iran, China, Malaysia, and a few other less represented countries. While no significant DIF indices were generated based on

nationality and age, biases were identified on some short-answer and multiple-choice items across ability groups.

Lee, Breland, and Muraki (2005) analyzed 81 TOEFL computer-based test (CBT) writing prompts based on the operational scores of over 250,000 test-takers from European or East Asian language backgrounds. A three-step logistic regression procedure for ordinal items was applied to the data and flagged about one third of the prompts for significant group effects. After examining the raw essay scores and the scores on the other sections of TOEFL, the researchers discussed that the essay score differences between the two language groups seemed to be similar to item impact (differences in ability) rather than a group difference attributable to a construct-irrelevant factor inherent in writing prompts. In other words, examinees of European language backgrounds would be expected to score higher on most TOEFL CBT writing prompts largely because they are of higher English language proficiency. Moreover, the effect sizes of the group differences were too small for any of the flagged prompts to be classified as having an important group effect.

Kim (2001) adopted both the likelihood ratio test and the logistic regression procedure in 1,038 students' scores on the *SPEAK* test (a retired speaking test developed by ETS) and found that the grammar and pronunciation subskills functioned differentially across European and the Asian language groups. Different from the content analysis conducted in Lee et al.'s (2005) study, which looked at writing prompts only (summarized in Breland et al,. 2004), Kim's study examined both the items and the scoring rubric and reached the conclusion that the types and the numbers of scoring scales might influence the validity of the test.

Differential Item Functioning in Locally Developed English Proficiency/Placement Tests

Some locally developed English tests have also adopted DIF for validation purposes. Such tests include the ESL Placement Exam at University of California, Los Angeles (Chen & Henning, 1985; Kunnan, 1990; Sasaki, 1991), the English Proficiency Test in China (Lin & Wu, 2003), the English subtest of the Korean National Entrance Exam (Pae, 2004), the English Placement Test in Japan (Shimizu & Zumbo, 2005), and an L2 vocabulary test in Finland (Takala & Kaftandjieva, 2000). As these studies either had smaller sample sizes (e.g., the UCLA ESL Placement Exam) or used grouping variables such as gender, major, and so forth (because they are administered to students from homogeneous native language backgrounds), they are not as relevant to our current study.

Meta-Analysis

Ferne and Rupp (2007) reviewed 27 DIF studies in language testing between 1990 and 2005. The studies reviewed included investigations on tests of English as well as other languages and tests of both first and second/foreign languages. The studies were compared on five essential sets of features: (a) the tests utilized, (b) the learner groups employed, (c) the DIF detection methods applied, (d) the reporting of DIF effects, and (e) the explanations for and consequences drawn from DIF results. The researchers critically pointed out the heterogeneity in methods and a general lack of explanatory power of predictor variables that made the 15 years of DIF research in language testing unreliable for future test construction. Detailed recommendations were provided on the descriptions of tests and learner characteristics, the adoption of specific hypotheses about the expected DIF effects from a confirmatory perspective, and the match between DIF methods and the structure (in test layout or scoring) of the language assessments being analyzed.

This Research Study

When any new assessment program initiates operational administration, it is critical to conduct validity studies to ensure that the test forms, items, and tasks associated with this assessment program are functioning as intended. In the case of TOEFL Junior Standard, several research studies were commissioned to investigate certain aspects of the assessments. This study focused on two specific psychometric characteristics of TOEFL Junior Standard:

1. Which segments of the assessment's score scales provide sufficient information to support the use of the scores in potential placement decisions?
2. Do items exhibit significant DIF when comparisons are made between test-takers from different countries?

An understanding of these psychometric characteristics for TOEFL Junior Standard is critical for ensuring valid interpretations of the scores from this assessment program. The results and findings from this study can provide information that is useful for improving the assessments, where necessary. For this study, the two specific research questions were as follows:

1. Based on the test information functions generated using item response theory (IRT) models, for which segments of the score scale do TOEFL Junior Standard forms provide sufficient information to support placement decisions?

2. For the TOEFL Junior Standard form used in the October 2010 administration, which items exhibited significant DIF by country? Are there linguistic features of these items that account for these DIF results?

It is important to examine the item statistics (ISs) from operational administrations to evaluate whether the items and tasks have a sufficient range of difficulty values to support use in potential placement decisions. TOEFL Junior Standard scores have already mapped onto the Common European Framework of Reference for Languages (CEFR). For Listening Comprehension, scaled scores of 225, 250, and 290 are the lowest scores corresponding to CEFR levels of A2, B1, and B2, respectively, while for Reading Comprehension, scaled scores of 210, 245, and 280 are the lowest scores corresponding to CEFR levels of A2, B1, and B2, respectively.

Methods

For this study, we used data from the first two operational administrations of TOEFL Junior Standard (October 2010 and February 2011). For each testing date, only one TOEFL Junior Standard form was administered (4GTJP01A in October 2010 and 4GTJP02A in February 2011). Of the 42 items in each section of the test, 30 were operational items and the other 12 were items either being pretested or serving as common items to be used in equating. The two TOEFL Junior Standard forms do not have any items in common. IRT statistics for all test items and test forms were computed based on the two-parameter logistic (2-PL) model using PARSCALE. We used the 2-PL model in this study because Sinharay, Haberman, and Jia (2011) compared the 2-PL and 3-PL models using TOEFL iBT data and found the 2-PL model preferable in terms of performance. Based upon their recommendation, TOEFL iBT has recently changed from a 3-PL model to a 2-PL model for operational work. In addition, the 2-PL model has better performance characteristics with smaller sample sizes. In the case of TOEFL Junior Standard, this may enable cross-country comparisons because the number of candidates per country per administration may be limited. Item information functions and test information functions were computed using ETS-developed programs. DIF analyses were conducted based on the standardization procedure developed by Dorans and Kulick (1986).

To determine whether items and tasks in each section have a sufficient range of difficulty and discrimination values, we first examined classical ISs (e.g., percent correct, point-biserial

correlation). We then calculated IRT statistics for the items, item information functions, and test information functions (TIFs) for each of the TOEFL Junior Standard test forms. The TIF, which is a simple sum of the information functions for items in a test, provides an overall evaluation of how much information a test is providing across the reporting scale (Hambleton & Lam, 2009). Analysis of the TIFs allowed us to determine the ranges of each score scale where it appears that the test information for that form is sufficiently high to support potential placement decisions. If the TIFs indicate that there is not enough information to support potential placement decisions for particular segments of the particular score scale, we could make recommendations as to the characteristics of additional items that are needed.

In addition, we conducted a series of DIF analyses by pairs of countries, limited to those where the sample sizes were sufficiently large (at least 500 test-takers per country per form). In the October 2010 administration, four countries (France, Greece, Korea, and Vietnam) met this sample size requirement. In the February 2011 administration, three countries (France, Greece, and Korea) met this sample size requirement.

Quantitative Results

Summary Statistics

In Table 1, we have reported the sample sizes by country by age of the test-takers for the October 2010 administration. In Tables 2 and 3, for the same test administration, we have reported the raw score means by country by age of the test-takers for each of the three sections of the test as well for the total composite score.

Table 1

Sample Sizes by Country by Age of Test-Takers

Age	Korea		France		Vietnam		Greece	
	Frequency	%	Frequency	%	Frequency	%	Frequency	%
11	-	-	-	-	-	-	22	4.94
12	178	13.03	-	-	-	-	161	36.18
13	660	48.32	-	-	-	-	178	40.00
14	427	31.26	75	9.66	248	13.42	84	18.88
15	101	7.39	299	38.53	1402	75.87	-	-
16	-	-	250	32.22	140	7.58	-	-
17	-	-	116	14.95	58	3.14	-	-
18	-	-	36	4.64	-	-	-	-
Total	1366	100.00	776	100.00	1848	100.00	445	100.00
Mean	13.33		15.66		15.00		12.73	
SD	0.79		1.00		0.58		0.82	

Table 2

Raw Score Means by Country by Age of Test-Takers: Listening Comprehension and Language Forms and Meaning Sections

Age	Korea			France			Vietnam			Greece		
	N	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD
Listening												
11	-	-	-	-	-	-	-	-	-	22	18.95	7.79
12	178	25.60	8.86	-	-	-	-	-	-	161	17.89	6.22
13	660	26.02	8.67	-	-	-	-	-	-	178	20.01	7.02
14	427	26.94	8.04	75	24.03	7.51	248	23.04	9.31	84	20.20	7.06
15	101	26.96	9.34	299	22.48	7.81	1,402	22.45	9.01	-	-	-
16	-	-	-	250	22.93	8.05	140	24.42	9.62	-	-	-
17	-	-	-	116	20.84	7.27	58	25.29	8.44	-	-	-
18	-	-	-	36	20.83	7.02	-	-	-	-	-	-
Total	1,366	26.32	8.56	776	22.45	7.78	1,848	22.77	9.10	445	19.23	6.85
LFM												
11	-	-	-	-	-	-	-	-	-	22	20.32	7.63
12	178	23.92	8.38	-	-	-	-	-	-	161	18.36	6.96
13	660	24.40	8.28	-	-	-	-	-	-	178	20.01	7.83
14	427	25.83	7.67	75	26.11	7.34	248	27.94	8.76	84	20.42	8.07
15	101	27.09	8.21	299	23.90	7.79	1402	28.15	8.60	-	-	-
16	-	-	-	250	24.22	8.70	140	26.80	8.60	-	-	-
17	-	-	-	116	22.71	7.95	58	26.52	8.74	-	-	-
18	-	-	-	36	22.67	7.24	-	-	-	-	-	-
Total	1,366	24.99	8.15	776	23.98	8.08	1,848	27.97	8.63	445	19.50	7.59

Note. LFM = Language Form and Meaning.

Table 3

Raw Score Means by Country by Age of Test-Takers: Reading Comprehension Section and Total Composite Raw Scores

Age	Korea			France			Vietnam			Greece		
	N	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD
Reading												
11	-	-	-	-	-	-	-	-	-	22	18.91	9.34
12	178	24.42	9.32	-	-	-	-	-	-	161	16.94	6.65
13	660	25.12	9.18	-	-	-	-	-	-	178	18.70	7.63
14	427	26.94	8.57	75	27.44	8.72	248	28.49	8.67	84	18.86	7.57
15	101	27.65	9.23	299	26.19	8.68	1,402	28.61	8.65	-	-	-
16	-	-	-	250	26.89	8.78	140	28.46	8.40	-	-	-
17	-	-	-	116	26.16	7.62	58	27.29	9.27	-	-	-
18	-	-	-	36	26.14	7.19	-	-	-	-	-	-
Total	1,366	25.79	9.07	776	26.53	8.50	1,848	28.54	8.65	445	18.10	7.40

Age	Korea			France			Vietnam			Greece		
	N	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD
						Composite						
11	-	-	-	-	-	-	-	-	-	22	58.18	22.58
12	178	73.94	25.29	-	-	-	-	-	-	161	53.19	17.46
13	660	75.54	24.63	-	-	-	-	-	-	178	58.71	20.55
14	427	79.71	22.66	75	77.57	21.47	248	79.47	25.11	84	59.48	21.05
15	101	81.70	25.72	299	72.57	22.60	1,402	79.21	24.50	-	-	-
16	-	-	-	250	74.04	23.80	140	79.69	25.25	-	-	-
17	-	-	-	116	69.71	21.22	58	79.10	23.79	-	-	-
18	-	-	-	36	69.64	19.78	-	-	-	-	-	-
Total	1,366	77.09	24.31	776	72.97	22.62	1848	79.28	24.60	445	56.83	19.82

Note. LFM = Language Form and Meaning.

Test Information Analyses

To answer the first research question, we analyzed the TIFs generated using the 2-PL model for each of the two forms of TOEFL Junior Standard. As a criterion for evaluating the TOEFL Junior Standard test forms from an IRT perspective, we chose two values for the TIF, 5 and 10, as standards by which to judge the adequacy of test information. A TIF value of 5 is equivalent to a classical test theory reliability estimate of .80, while a TIF value of 10 is equivalent to a classical test theory reliability estimate of .90 (Hambleton & Lam, 2009). In Figures 1–3, we display the TIFs for both forms by test section (Listening Comprehension; Language Form and Meaning; Reading Comprehension).

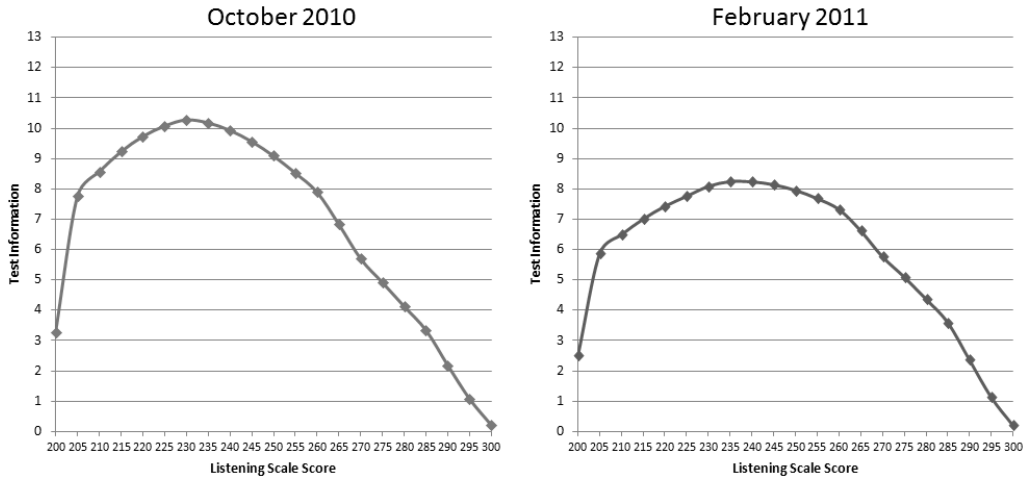

Figure 1. **Test information functions for Listening—October 2010 and February 2011 forms.**

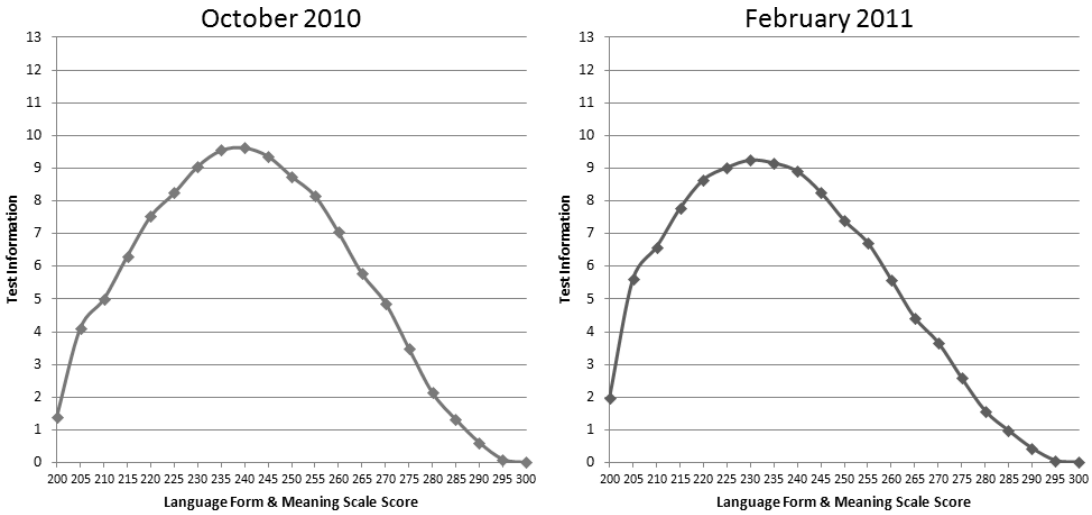

Figure 2. **Test information functions for Language Form and Meaning—October 2010 and February 2011 forms.**

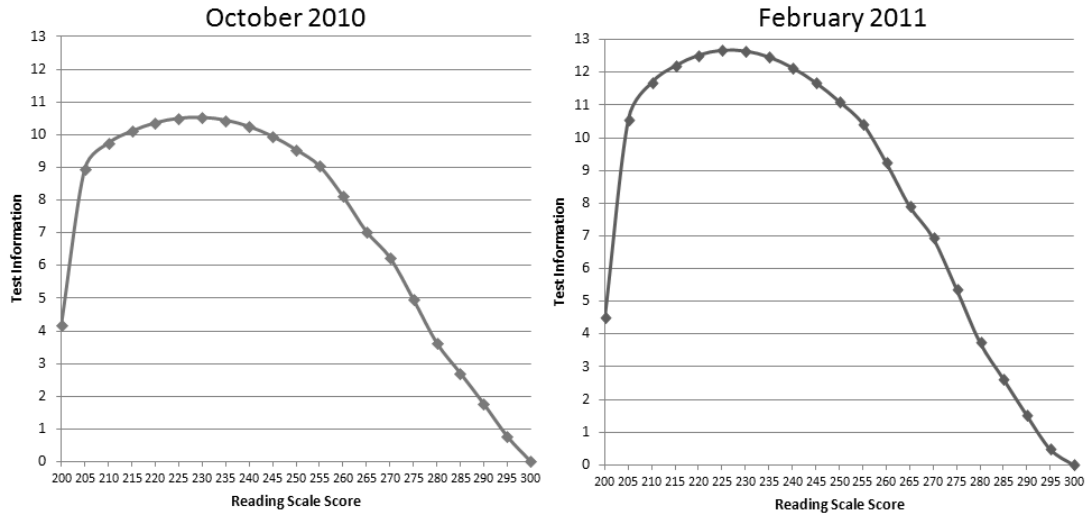

Figure 3. **Test information functions for Reading—October 2010 and February 2011 forms.**

Listening Comprehension

For the Listening Comprehension section, the TIF for the form used in the October 2010 administration shows that, for a range of scale scores from 200 to 276, the TIF for this section is greater than 5, and for a range of scale scores from 228 to 242, the TIF is greater than 10 (as a reminder, the score scale range for each section is from 200 to 300). For the form used in the February 2011 administration, the TIF for the Listening Comprehension section shows that, for a range of scale scores from 201 to 280, the TIF for this section is greater than 5. For this form, the

TIF did not exceed 10 on any portion of the score scale. Thus, for both TOEFL Junior Standard forms, the TIF for the Listening Comprehension section indicates a moderate level of information about test-takers across most of the score scale.

We then compared the TIFs for the Listening Comprehension section to obtain the Relative Efficiency of the two test forms (see Figure 4). The main purpose of providing information on the relative efficiency of the two forms is to indicate the variability in the degree of information provided by the forms across the score scale. In general, test forms differ to the extent to which they provide a high degree of test information. Although an examinee may only take a specific test form, it is useful for decision-makers to understand which test form provides a greater amount of information for different points on the score scale.

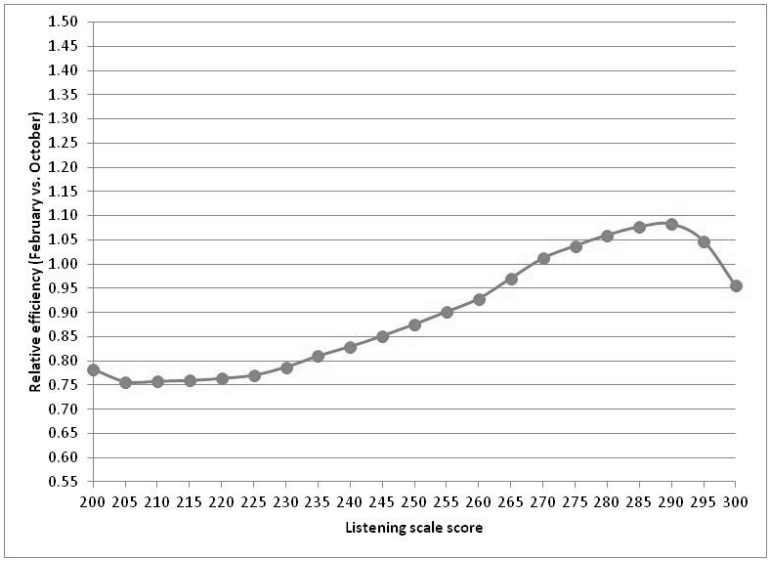

Figure 4. **Relative efficiency graph for Listening—February 2011 vs. October 2010 forms.**

Relative efficiency is computed as the ratio of the TIF for two test forms across the entire score scale. Either test form can be chosen as the base form; we chose the October 2010 form as the base form. Relative efficiency values greater than one indicate that the February 2011 form provides more information on test-takers while relative efficiency values less than one indicate that the October 2010 form is more informative. For the large majority of the score scale, from 200 to 272, the October 2010 form provides more information than the February 2011 form. The October 2010 form is relatively more informative at the low end of the score scale, particularly

for scores below 238 where relative efficiency is 0.80 or below. However, for scaled scores above 273 on the Listening Comprehension section, the February 2011 form is more informative.

Language Form and Meaning

For the Language Form and Meaning section, the TIF for the form used in the October 2010 administration shows that, for a range of scaled scores from approximately 205 to 260, the TIF for this section is greater than 5. For the form used in the February 2011 administration, the TIF shows that, for a range of scaled scores from 200 to 255, the TIF for this section is greater than 5. The TIFs for both forms did not exceed 10 on any portion of the score scale. Thus, for both TOEFL Junior Standard forms, the TIF for the Language Form and Meaning section indicates a moderate level of information about test-takers for scale scores of 260 or less.

We compared the TIFs for the Language Form and Meaning section to obtain the relative efficiency of the two test forms (see Figure 5). Again, we chose the October 2010 form as the base form. For the large majority of the score scale, from 222 to 300, the October 2010 form provides more information than the February 2011 form. For scores below 222 on the Language Form and Meaning section, the February 2011 form is more informative. The October 2010 form is relatively more informative at the high end of the score scale, particularly for scaled scores above 255.

Reading Comprehension

For the Reading Comprehension sections of both test forms, the analyses shows that, for the range of scaled scores from 200 to 268, the TIF for this section is greater than 5. For the October 2010 form, the TIF for this section is greater than 10 for scale scores from 214 to 237. For the form used in the February 2011 administration, the TIF for this section is greater than 10 for scale scores from 204 to 250. For both TOEFL Junior Standard forms, the TIF for the Reading Comprehension section indicates a moderate level of information about test-takers across most of the score scale (268 or below) and a higher level of information for scale scores from 214 to 237.

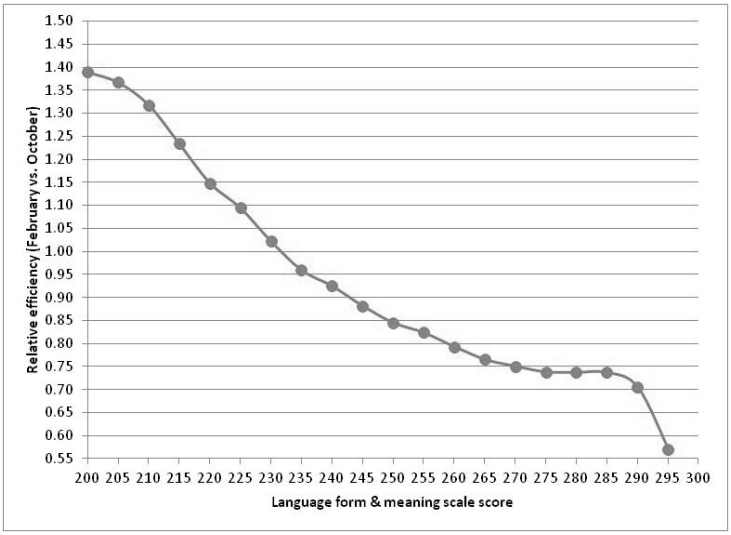

Figure 5. **Relative efficiency graph for Language Form and Meaning—February 2011 vs. October 2010 forms.**

We also compared the TIFs for the Reading Comprehension section to obtain the Relative Efficiency of the two test forms (see Figure 6). As with the other two sections, we chose the October 2010 form as the base form. For the large majority of the score scale, from 200 to about 276, the February 2011 form provides more information than the October 2010 form. Particularly, for scaled scores between 210 and 228 on the Reading Comprehension section, the February 2011 form is clearly more informative. The October 2010 form is relatively more informative only at the very high end of the score scale, for scale scores above 276.

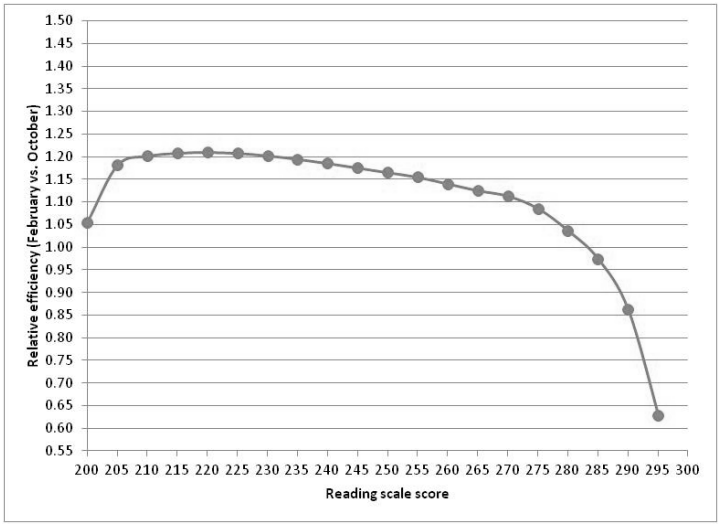

Figure 6. **Relative efficiency graph for Reading—February 2011 vs. October 2010 forms.**

With regard to the CEFR score boundaries, the most informative measurement was found for the A2 CEFR level (scaled scores of 225 to 250 for Listening Comprehension and 210 to 245 on Reading Comprehension). The least informative measurement occurred in the B2 CEFR level (scaled scores of 290 and higher on both Listening Comprehension and Reading Comprehension).

Differential Item Functioning Analyses

To investigate incidences of DIF, we conducted pairwise analyses of countries with at least 500 test-takers. These analyses were conducted separately for each form of TOEFL Junior Standard. For the October 2010 administration, there were four countries (France, Greece, Korea, and Vietnam) that met the sample size requirement, which produced six pairwise comparisons. For the February 2011 administration, there were three countries (France, Greece, and Korea) that met the sample size requirement, which produced three pairwise comparisons.

We focused on those pairwise comparisons that yielded evidence of significant (or C-level) DIF (Zieky, 1993). Tables 4 and 5 summarize the pairwise comparisons that showed C-level DIF for each item.

Table 4

Counts of C-Level DIF Items in the Pairwise Country Comparisons by Section—October 2010 Administration

		Listening Comprehension			Language Form and Meaning			Reading Comprehension		
Reference group	Focal group	Total	R	F	Total	R	F	Total	R	F
Korea	France	2	1	1	10	7	3	14	7	7
Korea	Greece	1	1	0	3	0	3	2	1	1
Korea	Vietnam	5	3	2	7	3	4	5	3	2
France	Greece	2	2	0	7	6	1	8	4	4
Vietnam	France	4	2	2	11	7	4	10	4	6
Vietnam	Greece	1	1	0	5	3	2	3	2	1

Note. DIF = differential item functioning; R = reference group; F = focal group.

Table 5

Counts of C-Level DIF Items in the Pairwise Country Comparisons by Section—February 2011 Administration

		Listening Comprehension			Language Form and Meaning			Reading Comprehension		
Reference group	Focal group	Total	R	F	Total	R	F	Total	R	F
Korea	France	7	2	5	10	7	3	9	4	5
Korea	Greece	4	3	1	6	4	2	9	5	4
France	Greece	1	1	0	6	3	3	4	2	2

Note. DIF = differential item functioning; R = reference group; F = focal group.

In addition, Tables 6 and 7 summarize the pairwise comparisons that showed moderate (or B-level DIF) for each item. Because of the large number of items that showed C-level DIF, we have confined our analyses to those items and have not included the ones that showed B-level DIF. The results indicated that, for both forms, there were fewer comparisons that showed evidence of C-level DIF in the Listening Comprehension section than in the Language Form and Meaning section or the Reading Comprehension section. In addition, some of the pairwise comparisons led to a greater number of items that exhibited C-level DIF. As supplementary analyses to the pairwise comparisons by countries, we also undertook an investigation of the linguistic features of items that exhibited DIF in order to develop a qualitative understanding of the features that can try to account for the DIF results. The results from these analyses are reported in the following section.

Table 6

Counts of B-Level DIF Items in the Pairwise Country Comparisons by Section—October 2010 Administration

		Listening Comprehension			Language Form and Meaning			Reading Comprehension		
Reference group	Focal group	Total	R	F	Total	R	F	Total	R	F
Korea	France	5	3	2	12	4	8	3	2	1
Korea	Greece	6	2	4	8	6	2	8	5	3
Korea	Vietnam	4	2	2	5	3	2	3	1	2
France	Greece	2	2	0	8	5	3	6	3	3
Vietnam	France	1	1	0	5	1	4	5	3	2
Vietnam	Greece	4	2	2	4	1	3	2	1	1

Note. DIF = differential item functioning; R = reference group; F = focal group.

Table 7

Counts of B-Level DIF Items in the Pairwise Country Comparisons by Section–February 2011 Administration

		Listening Comprehension			Language Form and Meaning			Reading Comprehension		
Reference group	Focal group	Total	R	F	Total	R	F	Total	R	F
Korea	France	5	5	0	4	0	4	6	4	2
Korea	Greece	4	2	2	4	3	1	3	1	2
France	Greece	4	1	3	4	2	2	5	2	3

Note. DIF = differential item functioning; R = reference group; F = focal group.

Qualitative Results

Linguistic Analyses of Differential Item Functioning Items

In order to develop a better understanding of the results from the DIF analyses, we undertook an investigation of the linguistic features of these items. The linguistic analyses were undertaken only for the October 2010 form. All of the items from this form are now retired and will not be used in future administrations, while the February 2011 form will continue to be used operationally. The research team for this project included the lead assessment developer for TOEFL Junior Standard, who participated in this investigation. The goal of this linguistic analysis was to determine whether the DIF findings between pairs of countries could be explained by our knowledge of differences across countries in one of the following:

- The linguistic differences or similarities between the predominant language spoken in the two countries and English
- The approaches or methods that are commonly used to teach English in the two countries
- Age- or grade-level differences in the test-taker samples from the two countries

As with all post-hoc analyses of DIF results, the findings we have reported here should be considered extremely preliminary, because we are attempting to infer, after the fact, possible causes behind DIF results. Because we do not have the insights that could be gained in an experimental study, one cannot and should not draw strong conclusions from the DIF results from this single observational study.

Consistent with our procedures for the DIF analyses, we focused on the pairwise country item level comparisons that exhibited C-level DIF, but only a particular subset of these items. An

IS represents the percent impact on item performance. A C-IS item is both C-DIF and has a percent correct difference between the matched groups of examinees (examinees from different countries) of 10% or more. Items that are not C-IS either did not reach the C-DIF threshold or did not have percent correct difference in the matched groups of 10% or more. Thus, C-IS items reach two problematic criteria (statistically reaching C-DIF status and being 10 or more percentage points harder for one group than the other, after matching the two groups on total test score). Thus, we limited our linguistic analyses to only those items with C-IS status (note that the number of DIF items discussed below differs, in some cases, from the numbers listed in Tables 4 and 5). The results from the linguistic analyses are organized first by section of the assessment, then by pairwise country analyses within each section.

Listening Comprehension

Korea–France comparisons. In these comparisons, two items exhibited C-level DIF, 2 and 31. Item 2 favored the Korean test-takers, while Item 31 favored the French test-takers. Item 2 tested lower level listening skills, specifically skills in recalling information. Item 31 tested higher level listening skills, specifically, being able to identify the implicit main idea of the listening passage. These two items also exhibited C-level DIF in the Korea–Vietnam comparisons, and the interpretation of these results is similar in so far as they may reflect age and/or grade-level differences between the test-takers from the two countries. That is, Item 2 favored the Korean test-takers because they are generally stronger in lower level listening skills, while Item 31 favored the French test-takers, who were older than the Korean test-takers and may have been more likely to possess the high-level cognitive skills needed to identify the implicit main idea of a passage.

Korea–Greece comparisons. In these comparisons, only one item exhibited C-level DIF, 11, which favored the Korean test-takers. Item 11 required test-takers to identify the implicit main idea of the listening passage. The Korean and Greek test-takers were similar in terms of ages and grade levels, so the DIF result may possibly be due to curriculum and/or instructional differences experienced by the students from the two countries.

Korea–Vietnam comparisons. In these comparisons, four items exhibited C-level DIF: 2, 31, 32, and 34. All but Item 2 came from the same stimulus passage. Two of these items, 2 and 32, favored the Korean test-takers, while the other two items, 31 and 34, favored the Vietnamese test-takers. Items 2 and 32 were both items that tested lower level listening skills; 2 was an item

that assessed skills in recalling information, while 32 required the recall of a phrase used in a sentence in the listening passage. Items 31 and 34 both tested higher level listening skills; Item 31 required test-takers to identify the implicit main idea of the passage, while Item 34 tested the ability to draw a weak inference. From what is known about the demographics and educational background of the students who participated, these results are consistent with the facts that the Korean test-takers are generally stronger in lower level listening skills as the form of instruction they received provides greater opportunities to develop oral skills more naturalistically.

In contrast, because the Vietnamese test-takers were generally older that the Korean test-takers (most of the Vietnamese test-takers were high school students while the Korean test-takers were predominantly middle school students), the Vietnamese students presumably possessed the high-level cognitive skills needed to identify the implicit main idea of a passage or to draw an inference more readily than the younger Korean students. Note, however, that this interpretation should be considered as speculative because Item 31 was one of the three items in this section that assesses a student's ability to identify the implicit main idea of the passage and the other two items did not exhibit DIF in the comparisons between Korea and Vietnam. Similarly, Item 34 was one of the seven items in this section that assessed a student's ability to draw a weak inference and the other six items did not exhibit DIF in the comparisons between these two countries.

France–Greece comparisons. In these comparisons, Items 11 and 12 exhibited C-level DIF and both items, which were linked to the same listening passage, favored the French test-takers. Item 11 required test-takers to identify the implicit main idea of the passage, while Item 12 was a question about a detail given in the passage and the correct response is a paraphrase of what was stated explicitly. For these two items, the DIF results may be due to age/grade-level and/or ability differences between test-takers from these two countries, as the French students were older and generally scored higher than the Greek students. The French students presumably possessed the high-level cognitive skills needed to identify the implicit main idea of a passage or to identify the correct paraphrasing of a detail more readily than the younger Greek students did.

Vietnam–France comparisons. In these comparisons, three items exhibited C-level DIF: 3, 12, and 34. Two of these items, 3 and 34, favored the Vietnamese test-takers, while Item 12 favored the French test-takers. Items 3 and 34 both tested higher level inferential skills, while Item 12 was a question about a detail given in the passage and the correct response is a

paraphrase of what was stated explicitly. The French and Vietnamese students are similar in ages and grade levels, so the DIF results are possibly due to curriculum and/or instructional differences experienced by students from the two countries.

Vietnam–Greece comparisons. In these comparisons, only one item exhibited C-level DIF, 39, which favored the Vietnamese test-takers. This item tested the main idea of an academic talk (the stem asks, What question is answered by the talk?), but it was unclear as to why this item favored the Vietnamese test-takers.

Language Form and Meaning

Korea–France comparisons. In these comparisons, 10 items exhibited C-level DIF: 4, 7, 8, 14, 20, 22, 24, 37, 41, and 42. Seven of the items, 4, 7, 8, 14, 20, 37, and 41, favored the Korean test-takers, while three items, 22, 24, and 42, favored the French test-takers. Items 20, 22, and 24 were linked to a stimulus article from a magazine about modern technology, while Items 37, 41, and 42 were associated with a reading passage from a history textbook. Two of the items that favored the French test-takers, 22 and 42, were items that tested vocabulary usage. For these items, all of the response choices had their roots in Latin, which tended to favor the French students, because French and English share many cognates.

Korea–Greece comparisons. In these comparisons, three items exhibited C-level DIF, 6, 13 and 24, all of which favored the Greek test-takers. All of these items tested straightforward grammatical usage and may have favored the Greek students due to curriculum and/or instructional differences experienced by the students from the two countries.

Korea–Vietnam comparisons. In these comparisons, six items exhibited C-level DIF: 24, 26, 36, 38, 40, and 42. Items 24 and 26 are linked to a stimulus article from a magazine about modern technology, while Items 36, 38, 40, and 42 are associated with a reading passage from a history textbook. Three of the items, 26, 40, and 42, favored the Korean test-takers, while the other three items, 24, 36, and 38, favored the Vietnamese test-takers. The items that favored the Korean test-takers were ones that, in terms of form, were more similar to an oral, rather than a written, construction. For these items, the DIF results are hypothesized to be due to the greater use of naturalistic, oral instruction in school that the Korean test-takers typically experience. The private schools in Korea as well as the Chungdahm afterschool enrichment programs, whose students make up the large majority of the Korean test-takers on which our report is based, teach English in a near-immersion environment where most teachers are native or near-native

speakers. Therefore, these students will certainly develop higher aural skills than students learning English in a traditional English-as-a-foreign-language model in which the native language might be used at times in instruction. We do not have evidence of how students are being taught in Vietnam; however, we do know that the Vietnamese students were older on average, while the Korean students were much younger on average. In addition, the range of vocabulary is generally greater for the Korean students, which helps in answering these items. In contrast, the items that favored the Vietnamese test-takers were based on passive structures, which are generally easier for older students to understand.

France–Greece comparisons. In these comparisons, six items exhibited C-level DIF, 2, 13, 18, 19, 22, and 42. Five of these items favored the French test-takers; the only item that favored the Greek test-takers was 13. Item 13 has a key of *often*, with an attractive distractor of *frequent*. This could be a case of the French test-takers being fooled by the cognate *frequent*, which in this case is not the correct form, because to be correct, it would have to be followed by *-ly*. The Greek test-takers may not have relied on cognates in their learning and would have learned the word *often* and recognized it as the proper choice. For all of the other items, it made sense that the French test-takers would be favored because French is generally more similar to English than Greek is.

Vietnam–France comparisons. In these comparisons, 10 items exhibited C-level DIF: 4, 8, 13, 14, 17, 22, 35, 37, 38, and 42. Four of these items, 13, 22, 36, and 42, favored the French test-takers, while six items, 4, 8, 14, 17, 37, and 38, favored the Vietnamese test-takers. Item 13 is the *often/frequent* item discussed above. Because Vietnam was once a French colony and once had a high level of exposure to the French language, the Vietnamese test-takers may have been fooled in the same way as the French students. Item 22 was a vocabulary item with direct cognates; Item 36 tested a relative clause, a structure that also exists in French; and 42 was another vocabulary item based on cognates. The other six items tested varying grammar points, but because the Vietnamese test-takers were generally higher performing on the test overall, perhaps these are points they may have learned better.

Vietnam–Greece comparisons. In these comparisons, five items exhibited C-level DIF: 2, 6, 13, 14, and 17. Three of these items, 2, 14, and 17, favored the Vietnamese test-takers, while two items, 6 and 13, favored the Greek test-takers. Item 13 is the *often/frequent* item that the Greeks seemed to have learned well. Item 6 tests the expletive *It is* (as in “It is always best”);

this may be a structure the Greek students learned that is not so familiar to the Vietnamese test-takers. Because the Vietnamese test-takers were generally better performing overall than the Greek test-takers, we hypothesized that they are better schooled in the points tested in the other items.

Reading Comprehension

Korea–France comparisons. In these comparisons, 12 items exhibited C-level DIF, 2, 9, 13, 14, 15, 22, 25, 26, 35, 38, 40, and 42. Five of the items, 15, 22, 35, 38, and 42, favored the Korean test-takers, while seven items, 2, 9, 13, 14, 25, 26, and 40, favored the French test-takers. There are several possible explanations for why the Korean test-takers performed better on the items that exhibited DIF in their favor. On average, the French test-takers performed slightly better than the Korean students on the Reading Comprehension section of TOEFL Junior Standard. Items 22 and 35 both required students to make a higher level inference from their reading, a task in which the Korean test-takers may have been stronger than the French students. Item 38 asked about the meaning of a phrase that does not have a Latinate root; students were asked to determine the meaning of the phrase from the context of the reading passage. Item 42 required test-takers to make an inference based on an indirect phrasing; again, this may have been a task in which the Korean test-takers may have been stronger than the French students.

Items 2 and 9 were items that were cognitively more demanding, favored the French test-takers who were older than the Korean test-takers and were more likely to possess the high-level cognitive skills needed to identify the implicit main idea of a passage. Item 13 asked about the meaning of a word, *cuisine*, that has French and Latin roots. Item 14 asked about how students will *show their appreciation*, an idiomatic expression that uses a word with French and Latin roots. Item 25 asked about the meaning of a word, *erect*, that has a Latin cognate. Item 40 asked students to identify the meaning of a pronoun, and the correct answer, substance, is a high-level vocabulary word that is more likely to be known to the French test-takers, who were older than the Korean test-takers.

Korea–Greece comparisons. In these comparisons, two items exhibited C-level DIF, 13 and 15. Item 15 favored the Korean test-takers, while 13 favored the Greek test-takers. Item 13 tests *cuisine*, which, while French in origin, is perhaps a pan-European word that would be more familiar to the Greek students than to the Korean studen

Korea–Vietnam comparisons. In these comparisons, four items exhibited C-level DIF: 12, 15, 22, and 35. Two items, 15 and 35, favored the Korean test-takers, while the other two items, 12 and 22, favored the Vietnamese test-takers. Item 12 favors the Vietnamese students and tests a direct French cognate, *venue*, perhaps a legacy of the imprint of French colonialism on Vietnamese culture.

France–Greece comparisons. In these comparisons, eight items exhibited C-level DIF, 2, 4, 9, 12, 17, 19, 40, and 42. Four of these items, 2, 4, 9, and 40, favored the French test-takers, while four items favored the Greek test-takers, 12, 17, 19, and 42. For Item 2, multiple steps were required to develop an inference to correctly answer this question; this process may have favored the French test-takers, who were generally older than the Greek test-takers. Similarly, Item 4 required students to carefully read the notes associated with a time schedule, a task which may have been easier for the older French test-takers. Item 9 asked students to identify the best headline for a newspaper article, an inference task that required high-level comprehension skills that were more likely to be possessed by the older French students. Item 40 asked students to identify the meaning of a pronoun, and the correct answer, *substance*, is a high-level vocabulary word that is more likely to be known to the French test-takers.

Of the items that favored the Greek test-takers, 12 inquired about the meaning of a word, *venue.* For this item, knowledge of the Latinate root of this word may have led the French test-takers to choose an incorrect option. Item 42 required test-takers to make an inference based on a direct phrasing; this may have been a relatively easy task for the Greek students.

Vietnam–France comparisons. In these comparisons, eight items exhibited C-level DIF: 2, 12, 13, 14, 20, 22, 41, and 42. Four of these items, 2, 13, 14, and 41, favored the French test-takers, while the other four items, 12, 20, 22, and 42, favored the Vietnamese test-takers. For Item 2, multiple steps were required to develop an inference to correctly answer this question; this process may have favored the French test-takers, who were slightly older than the Vietnamese test-takers. The results for Items 13 and 14 are similar to those found in the Korea–France comparisons. Item 13 asked about the meaning of a word, *cuisine*, that has French and Latin roots. Item 14 asked about how students will *show their appreciation,* an idiomatic expression that uses a word with French and Latin roots. Item 41 asked about the meaning of a word, *cuisine*, that has French and Latin roots.

Of the items that favored the Vietnamese test-takers, 12 inquired about the meaning of a word, *venue*. For this item, knowledge of the Latinate root of this word may have led the French test-takers to choose an incorrect option (similar to that found in the France–Greece comparisons). Item 20 required test-takers to make a prediction based on an inference regarding future action. Because the Vietnamese students were generally higher performing than the French test-takers, this item may have been unusually difficult for the French students. Item 42 required test-takers to make an inference based on a direct phrasing; this may have been a relatively easy task for the Vietnamese students.

Vietnam–Greece comparisons. In these comparisons, three items exhibited C-level DIF: 4, 13, and 22. Two items, 4 and 22, favored the Vietnamese test-takers, while Item 13 favored the Greek test-takers. Item 13 tests *cuisine*, which, while French in origin, is perhaps a pan-European word that would be more familiar to the Greek students than the Vietnamese students.

Discussion

In this study, two main research questions were addressed:

- Based on the test information functions generated using IRT models, for which segments of the score scale do TOEFL Junior Standard forms provide sufficient information to support placement decisions?
- For the TOEFL Junior Standard form used in the October 2010 administration, which items exhibit significant DIF by country? Are there linguistic features of these items that account for these DIF results?

For the first research question, both of the TOEFL Junior Standard forms we analyzed appear to provide sufficient information to support placement decisions across the majority of the score scale for all three sections of the assessment. More specifically, the TIFs indicate that there is, at a minimum, a moderate level of information provided about test-takers generally at scores of 280 or lower (the equivalent of the CEFR B2 level) for all three sections. The students in the A2 CEFR level are more effectively measured than the students in the B1 and B2 CEFR levels. There is a high level of information provided about test-takers for some parts of the score scale: For the Listening Comprehension section of the form used in the October 2010 administration, the TIF exceeded 10 for scale scores from 228 to 242. Additionally, for the Reading Comprehension section, the TIF exceeded 10 for scale scores from 214 to 237 for the

form used in the October 2010 administration as well as for scale scores from 204 to 250 for the form used in the February 2011 administration.

For the second research question, we found that a moderate number of items exhibited C-level DIF in the pairwise country comparisons. When compared with earlier investigations of DIF by country or language group in English language proficiency assessments, the proportion of items that exhibited significant DIF on these two forms of TOEFL Junior Standard was somewhat lower than was found in those prior studies. The linguistic analyses conducted on the DIF results showed plausible *construct-relevant* explanations for most of the findings. That is, for the large majority of the items that exhibited C-level DIF in the pairwise country comparisons, the results could be tied directly to at least one of three pre-existing group differences between the samples of test-takers across countries: (a) linguistic differences or similarities between the predominant language spoken in the two countries and English, (b) approaches or methods that are commonly used to teach English in the two countries, or (c) age or grade-level differences in the test-taker samples from the two countries. In addition to group differences at the item level, a review of the results at the item-type level indicated some interesting patterns:

- For the items that assessed vocabulary knowledge in the Language Form and Meaning and Reading Comprehension sections, the French test-takers appeared to perform about as well as the other test-takers. This may be due to the fact that, of the native languages in this study, French is the one morphologically closest to English.
- For the items that assessed inferential understanding in the Listening Comprehension and Reading Comprehension sections, the older test-takers appeared to perform better than the other test-takers on these more cognitively challenging items.

In summary, from a fairness perspective, the items in the TOEFL Junior Standard used in the October 2010 administration do not appear to be unfair to the test-takers from any of the countries because our linguistic analyses showed plausible construct-relevant explanations for most of the findings.

References

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language* (TOEFL Research Report No. 9). Princeton, NJ: Educational Testing Service.

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of the International English Language Testing System (IELTS) listening module. *International Journal of Listening, 26*(1), 40–60.

Aryadoust, V, Goh, C. C. M., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly, 8*(4), 361–385.

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. 76). Princeton, NJ: Educational Testing Service.

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

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.

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.

Hambleton, R. K., & Lam, W. (2009). *Redesign of MCAS tests based on a consideration of information functions* (MCAS Validity Report No. 18; CEA-689). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

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

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly*, *24,* 741–744.

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.

Lin, J., & Wu, F. (2003, April). *Differential performance by gender in foreign language testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Liu, O. L., Schedl, M., Malloy, J., & Kong, N. (2009). *Does content knowledge affect TOEFL iBT™ reading performance? A confirmatory approach to differential item functioning* (Research Report No. RR-09-29). Princeton, NJ: Educational Testing Service.

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

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

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

Shimizu, Y., & Zumbo, B. D. (2005). Logistic regression for differential item functioning: A primer. *Japan Language Testing Association Journal, 7,* 110–124.

Sinharay, S., Haberman, S. J., & Jia, H. (2011). *Fit of item response theory models: A survey of data from several operational tests* (Research Report No. RR-11-29). Princeton, NJ: Educational Testing Service.

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

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

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