Listening. Learning. Leading.®

Research Report
ETS RR–14-09

Do *TOEFL iBT*® Scores Reflect Improvement in English-Language Proficiency? Extending the TOEFL iBT Validity Argument

Guangming Ling

Donald E. Powers

Rachel M. Adler

June 2014

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Research Scientist

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Managing Research Scientist

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Gary Ockey
Research Scientist

Donald Powers
Managing Principal Research Scientist

Gautam Puhan
Senior Psychometrician

John Sabatini
Managing Principal Research Scientist

Matthias von Davier
Director, Research

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Stellhorn
Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Do *TOEFL iBT*® Scores Reflect Improvement in English-Language Proficiency? Extending the TOEFL iBT Validity Argument

Guangming Ling, Donald E. Powers, & Rachel M. Adler

Educational Testing Service, Princeton, NJ

One fundamental way to determine the validity of standardized English-language test scores is to investigate the extent to which they reflect anticipated learning effects in different English-language programs. In this study, we investigated the extent to which the *TOEFL iBT*® practice test reflects the learning effects of students at intensive English programs in the United States and China, as well as extracurricular English-learning activities that may be associated with the expected learning effects. A total of 607 students at the high school level or beyond participated in the United States and China, including 111 students who took 2 forms of the practice test under a pretest and posttest design. The results showed moderate to substantial levels of improvement on each of the TOEFL iBT sections, with different score gain patterns for students in the United States and China. We concluded that students who study at English programs similar to those included in this study can improve their English-language proficiency levels at least moderately over 6 months or longer, as indicated by changes in their scores on the TOEFL iBT practice test. This improvement is consistent with an interpretation of TOEFL iBT scores as indicators of English-language proficiency.

Keywords Learning effect; intensive English program; extracurricular learning activity; weekly hours; TOEFL; language learning

doi:10.1002/ets2.12007

Validity refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores (Messick, 1989). The validation process must involve multiple means of accumulating evidence in support of a particular interpretation about test scores. For example, a validity argument can be supported based on whether, and to what extent, test scores reflect appropriate changes as a function of construct-related interventions, such as instruction or learning (Messick, 1989; see also Cronbach, 1971). In the context of English-language testing, it is important to examine the extent to which test scores capture the learning effects that may differ by English-language programs and the English learners themselves (Chapelle, Enright, & Jamieson, 2008). For a test like the *TOEFL iBT*® exam, such a validation process would require evidence of a test score increase following instruction and learning, and evidence demonstrating that the amount of score increase reflects individual differences among English learners and English-language programs (Chapelle et al., 2008, p. 20). Thus, the TOEFL iBT validity argument can be enhanced by obtaining a better understanding of the relationship between the characteristics of programs/individuals and the improvement in individuals' English-language proficiency, as captured by TOEFL iBT scores. This better understanding also has very practical utility for stakeholders, such as test takers, test score users, and educators, as they may be interested in knowing about whether, and how, these factors are related to test score improvements.

According to Ellis (1993, 2008), there are six general factors that may contribute to the acquisition of a second language: (a) the way that a second language is exposed to learners, such as via unidirectional input or interactions, (b) the social and situational factors related to language learning, (c) the degree to which language transfers between first and second language, (d) cognitive factors, (e) sociocultural factors, and (f) linguistic universals. Some of these factors are related to programs, such as curriculum design and instructional methods. Other factors pertain to individuals, such as students' motivation for learning English, learning style and strategy, time spent on English, and extracurricular English-learning activities. There seems unlikely to be any definitive study that could isolate the causes of improvement in English-language proficiency. Acknowledging this limitation, we conducted this study to validate the extent to which the TOEFL iBT scores

Corresponding author: G. Ling, E-mail: gling@ets.org

can reflect the learning effects for students at English-language programs in the United States and overseas. The learning effects are assumed to be observable and can be measured by increases on standardized test scores. We also expected that, with this study, we could obtain information about some factors that may be associated with assumed learning effects (e.g., score increase on an English-language test).

After an extensive literature search, we found only several studies that examined the factors that contribute to score improvements on major English tests, such as the *TOEFL*®, *TOEIC*®, and IELTS exams. We briefly summarize the results of these studies in a suitable way as to inform this study.

Several studies used the TOEFL exam to measure improvement in students' English proficiency as a result of formal instruction or specific interventions. Tanaka and Ellis (2003) examined changes in students' beliefs about English-language learning and their English-language proficiency (as reflected by TOEFL scores) over the course of a study abroad program. They found significant score gains on the TOEFL and significant changes in students' beliefs. However, the correlations between students' beliefs and their TOEFL scores were weak, which is inconsistent with other studies of the relationship between learners' beliefs and their test performance (Mori, 1999; Park, 1995). Wilson (1987) found that score changes were associated with test repetition status. Forster and Karn (1998) described strategies teachers could use to improve scores on the TOEIC and TOEFL tests, but they did not document changes that may have resulted from using the strategies.

Rea-Dickins and Scott (2007) introduced a special issue of *Assessment in Education* that dealt with the effects of different assessment mechanisms on learner performance (see also Ross, 2005). Other studies (Jiang, 2007) examined very specific instructional methods that improve students' English-language skills. Still others (Eggly & Schubiner, 1991) documented the effects of specific courses on particular language skills. A number of researchers have attempted to document strategies used by English-language learners studying independently, as opposed to studying for an English course, and have related the use of these strategies to improvements in language skills (Chin-Chin, 2007).

Several studies have documented learning effects using other major English tests. For example, Green (2007) found that taking classes dedicated to test preparation did not improve students' writing test scores on the IELTS when compared to taking regular academic writing classes or classes focusing on a combination of academic writing and test preparation. Elder and O'Loughlin (2003) found that the student's living environment (e.g., at home, in a family setting, or with fellow students in an intensive program), course level, educational qualifications, and reading proficiency together provided the best predictors of gains in overall IELTS scores, with a moderate relationship between nationality and score gains.

While most of the studies reviewed above can be categorized by one or more factors described by Ellis (2008), some of these factors focused in these studies, such as how the learning effects was measured, was not discussed in full details in Ellis's review. With the small number of similar studies, it is hard to reconcile the results and generalize, which indicates a need for further investigation. In addition, though some of these studies used a standardized test to measure learning effects, others used local or in-house measures of language proficiency, which may lead to ambiguous conclusions. This ambiguity may also be related to the fact that these studies were based on samples of English learners at different English-language programs, with varying levels of English-studying intensity and length, and from a wide range of social and environmental contexts. In addition, we found little research based in China or other Asian countries, even though the absolute number of English learners is large in these countries and they comprise a large proportion of the population of TOEFL, TOEIC, and IELTS test takers. Finally, no study was found to examine the impact of students' extracurricular activities on their English skills improvement as reflected on the score changes on standardized tests. It is desirable, therefore, to conduct further research in order to understand English learners' growth trajectories in these countries as well as those in the United States.

In this study, we focused on student factors that may be associated with English learning and score changes on the TOEFL iBT, as well as possible institutional differences associated with score changes. Acknowledging the challenges associated with measuring score changes in an operational setting, we decided to use a practice version of the TOEFL iBT (two test forms) in this study. It is recognized that students' practice test scores can only approximate their operational test scores, as many conditions in the practice test, including student motivation, differ from those in an operational test. Nevertheless, scores from the practice test were reasonably assumed to provide a suitable approximation of students' English proficiency levels as reflected in operational TOEFL iBT scores and could thus provide useful information to address the research questions identified in this study.

Research Questions

1 Do students improve their English-language proficiency over the course of the English-language programs, as seen on their score changes on the TOEFL iBT practice tests?
2 Are the score changes different among English programs? If so, are the differences associated with any of the program-related factors?
3 Are the score changes associated with students' background (e.g., gender) or with any of their self-reported extracurricular English-learning activities?

Method

Instruments

Two forms of the TOEFL iBT practice test were generated by using the TOEFL iBT Research Form Creator (see Ling & Bridgeman, 2013, for more details). The practice test used in this study resembled the operational TOEFL iBT in all respects, including the delivery platform and interface, with each form having four sections, including Reading, Listening, Speaking, and Writing.

A brief English-learning survey was administered following the posttest to collect information about students' background information such as gender, ethnicity, first language, education level, reasons to learn English, number of years and weekly extracurricular hours learning English, and experience with major English tests. The survey was administered in Chinese for students in China, and in English for students in the United States.

The survey also had a series of questions about (a) extracurricular English-learning activities (excluding homework), (b) the frequency with which students engaged in each of them, and (c) the extent to which any of the activities were acknowledged by students as effective ways to improve their English skills. A series of activities were listed, including reading English books, reading English magazines and newspapers, listening to or watching English media, participating in online discussions (e.g., blog, forum, text chatting, etc.), chatting through Internet with voice or video, reading aloud in English, participating in English salons or clubs, and practicing speaking skills with native English speakers. Students were asked to add any activities they engaged in but were missing in the survey.

Participants

A US-based intensive English program and an international high school in China participated in this study. The sample included 607 students, 480 from School A (in China) and 127 from School B (see Table 1). Among them, 111 students took the two practice tests (pretests and posttests) and were treated as the longitudinal sample to address the research questions directly (Table 1). Students were encouraged to participate in this study and take the tests to try authentic TOEFL iBT test items and to evaluate their performance based on the testing results. No monetary compensation was provided to students.

All the students from School A were native Chinese speakers, and half of them were males. At School B, there were more males than females (69 vs. 31%). These students had diverse first-language backgrounds, primarily Arabic, French,

Table 1 Participating Schools With Descriptions and Number of Students

School	Nature of English program	Description of English courses	*N* total	Time between tests	*N* tested twice
A	Chinese high school students	General English courses as part of K-12 English education required by government TOEFL iBT preparation courses	480	9 months	90
B	Mainly intensive	Intensive English courses only, covering reading, listening, speaking, and writing skills	127	6 months	21

Turkish, and Chinese. About one third of the students were at the high school level or below, another one third at the undergraduate level, and the remaining one third at the graduate level.

Among the participants at School A, 235 students took the posttest and answered the survey questions. About half of them (116) were finishing their first-year studies at the time of posttest, whereas the others (119) were finishing their second-year studies at that time. These students were treated as the cross-sectional sample in this study to address the research questions indirectly.

Design

Students took the first TOEFL iBT practice test (the pretest) at the beginning of the semester or school year and the second one (the posttest) at the end of the semester or school year. All the students took the two practice tests voluntarily but were not aware of each test until 1 or 2 weeks before the test date. The time between the pretest and posttest was different between the two schools, about 9 months at School A and 5 months at School B (Table 1).

Between the two tests, those students at School B took regular intensive English courses for 20 hours a week, but had no extra coaching or instructional courses directly targeted at the TOEFL iBT. At School A, students took the high school English classes required by the general educational guidelines in China between the two tests, together with classes on other subjects, such as math, Chinese language arts, physics, biology, and chemistry. In addition, they took TOEFL iBT preparation courses and exercises during the second half of the 9 months. The English-related course work was less than 15 hours a week on average. (see Table 1.)

Data

Data for each school came from two sources: students' section scores on the TOEFL iBT practice tests (pretest and posttest) and students' responses to the English-learning survey questions. All responses to the multiple choice items of Reading and Listening sections were scored using the scoring keys. The *e-rater*® scoring engine[1] was used to score the Writing section responses (essays). The *SpeechRater*^{SM} scoring engine[2] was used to score the Speaking section responses. There were cases where no score was produced because of limitations associated with low audio quality or low recognition rate by the SpeechRater. Speaking scores were treated as missing in the analyses for some students, even though they may have produced speaking responses. All section scores of the two forms were put on the same scale through a conversion table based on operational equating results such that the two sets of scores were comparable and were interpreted in the same way.

Analysis

Descriptive analyses were performed on the students' survey data and section scores, after being grouped by school, gender, and other background variables. General linear models (GLMs) were applied to examine whether students' section scores improved and whether the improvement was associated with students' learning activities. Cohen's effect size (d) was computed for each section score to examine whether the score gains were of any practical importance. Cohen's d is considered small for values between .2 and .3, moderate for values around .5, and large or substantial around .8 (Cohen, 1988, p. 25).

Results

Survey Results

A total of 300 students responded to the survey questions: 235 students from School A and 65 students from School B. All the 111 students who took both the pretest and the posttest responded to the survey questions and were also included in the analysis here.

As displayed in Figure 1, the reported reasons for learning English varied among students and schools. At School A, most (87%) students indicated that they were studying English to improve their English-language proficiency, half (51%) reported learning English because their parents or school required it, 86% of the students were studying in preparation

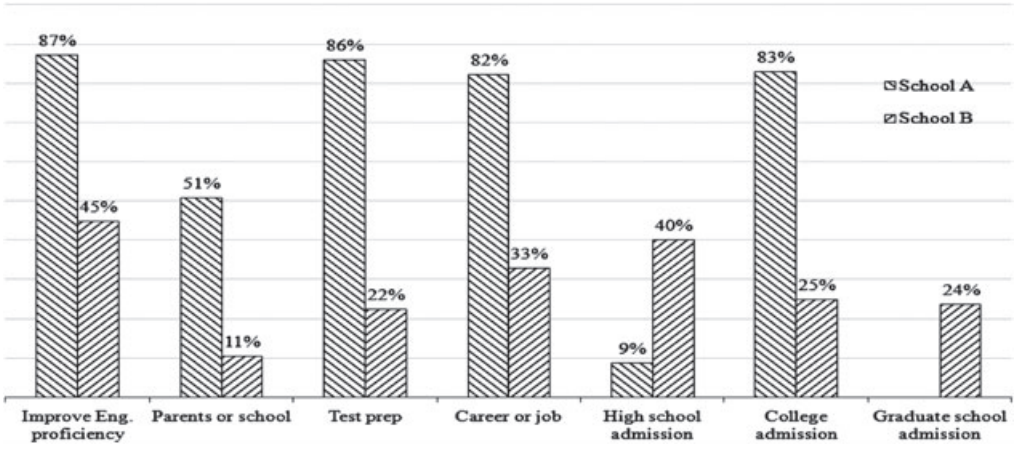

Figure 1 Reported reasons for learning English by school.

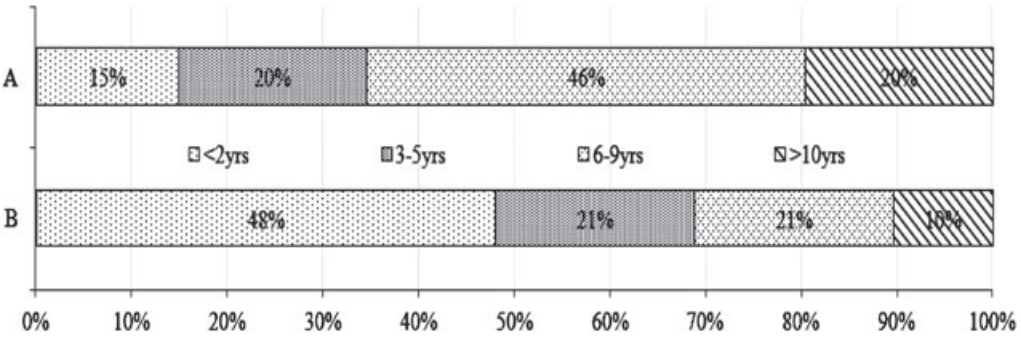

Figure 2 Reported number of years learning English by school.

for a language test, 82% felt that studying English would improve their job opportunities, and 83% of the students were studying English to gain college admission (Figure 1).

At School B, the percentage of students endorsing each category was much smaller than that at School A, with the lowest for students who reported learning English because it is required by their parents or school (11%) and the highest for students who reported studying English for future admission to an undergraduate (25%) or graduate program (24%) in an English-speaking country (totals up to 49%; Figure 1).

Half of the students at School A were second-year high school students and had taken the TOEFL iBT test prior to this study; the other half were first-year students who had not taken the test but were planning to do so. At School B, 66% of the students had taken the TOEFL iBT test, 36% had taken the IELTS test, and 9% had taken the TOEIC test.

The number of reported years of learning English also varied by student and school; 86% of the students at School A reported studying English for more than 3 years, and the majority of them (66%) said they studied English for more than 6 years (Figure 2). However, only half of the students (52%) at School B reported studying English for more than 3 years, and less than one third (31%) reported having studied English for more than 6 years.

Similarly, the reported total extracurricular time spent per week learning English varied by student and school. More than one third of the students (37%) at School A reported spending more than 10 hours a week studying English outside the classroom, and close to one fourth of the students (23%) reported that they spent 6–9 hours studying English. In contrast, the students at School B reported they spent fewer hours on average studying English outside the classroom, with 40% of them reporting studying less than 2 hours per week, and another 40% reporting studying 3–5 hours per week (Figure 3).

Two types of extracurricular activities for improving English reading skills appeared in the English-learning survey. A good portion of the students (more than 40%) reported reading English magazines or books on, at most, a monthly basis, regardless of the school they attended. More students at School B than at School A reported reading English books on a

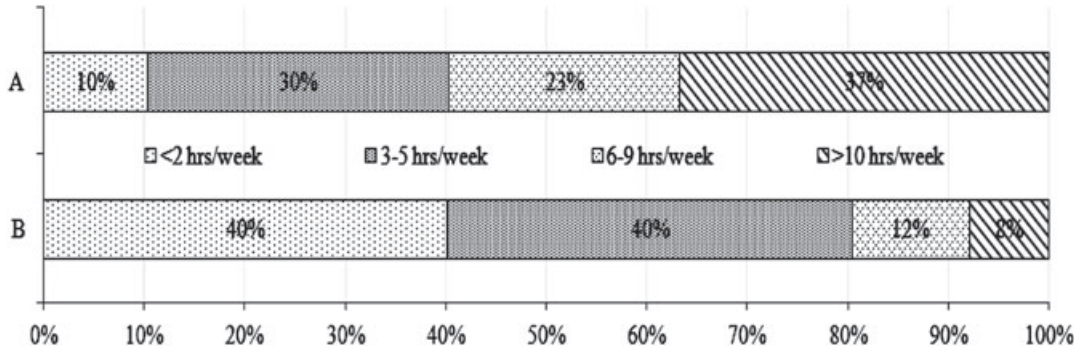

Figure 3 Reported number of extracurricular hours on learning English per week by school.

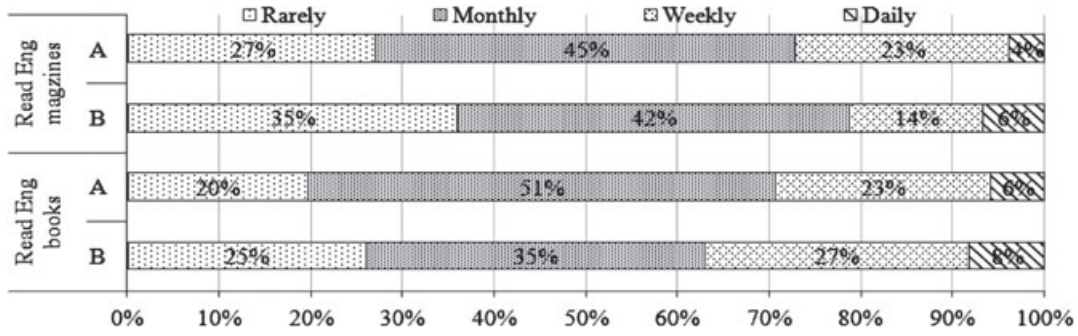

Figure 4 Reported English reading activities by school.

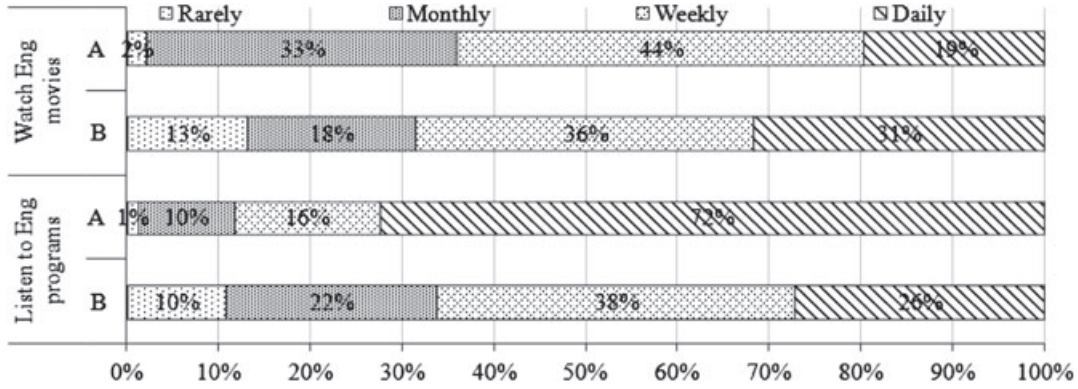

Figure 5 Reported English listening activities by school.

weekly or daily basis (35 vs. 29%). However, more students at School A (27%) than at School B (20%) reported reading English magazines or newspapers on a daily or weekly basis (Figure 4).

More than half of the students reported spending time listening to English programs (e.g., radio programs, and songs) or watching English movies on a daily or weekly basis, regardless of their schools. Most (72%) students at School A reported listening to English programs every day, whereas only one fourth (26%) did so at School B. The percentage of students reporting they spent time watching English movies on at least a weekly basis was comparable, 63% at School A and 67% at School B (Figure 5).

Figure 6 provides the results of reported writing-related extracurricular English-learning activities. Overall, students reported they spent less time on writing-related extracurricular English-learning activities than on listening-related activities; more than 40% of the students reported they spent time on writing-related activities on a monthly basis or even less frequently, regardless of school. More than half (56%) of the students at School B reported spending time writing text messages in English on phones or computers, comparing with 19% of students at School A. Similarly, more than half (55%)

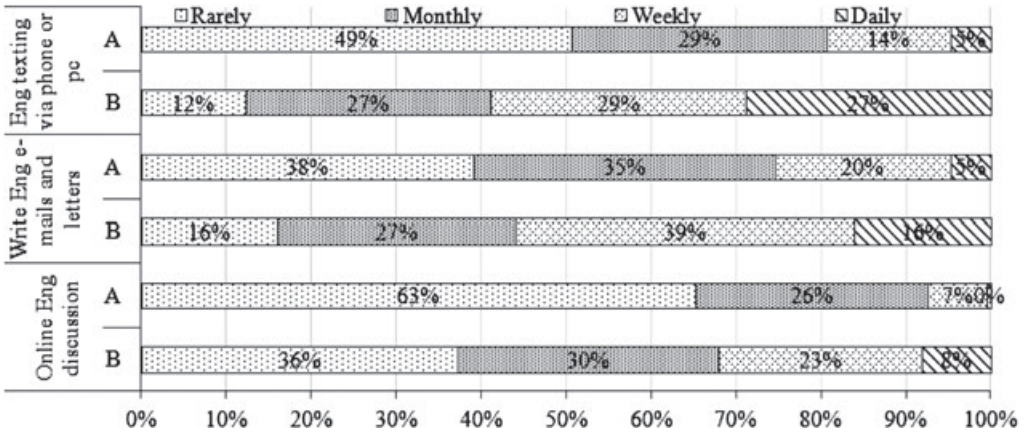

Figure 6 Reported English writing activities by school.

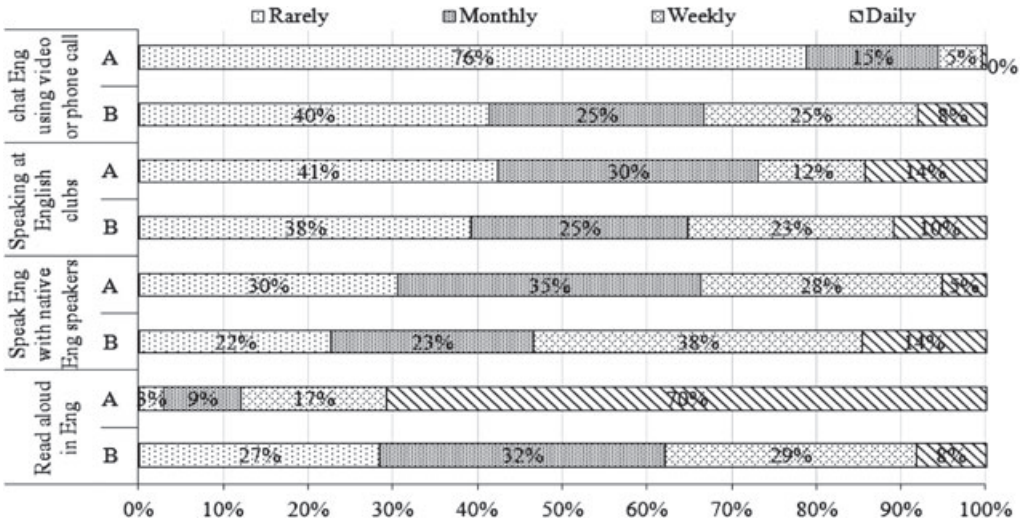

Figure 7 Reported English-speaking activities by school.

of the students at School B reported writing letters or e-mail messages in English on a daily or weekly basis, whereas only 25% did so at School A. Finally, 31% of the students at School B indicated that they participated in online discussions using English, whereas only 11% of the students at School A did so. It should be noted that, at School A, all students lived on campus and had limited or no access to cell phones and computers to engage in activities using social media.

Overall, students' engagement in speaking-related extracurricular activities were comparable to those in writing-related activities, as displayed in Figure 7. More than 40% of the students reported spending time on speaking-related activities on, at most, a monthly basis, regardless of school. About one third of students at School B (33%) said they speak English on video chats or phone calls on a daily or weekly basis, whereas this was the case for only 9% of the students at School A. More students at School B reported practicing speaking at English clubs on a daily or weekly basis than students at School A (33% and 26%, respectively). Similarly, a greater proportion of students at School B reported they conversed with a native English speaker on a daily or weekly basis than at School A (52% and 33%, respectively). However, 70% of the students at School A reported participating in read-aloud exercises on a daily basis, whereas only 8% of the students at School B reported doing so.

Overall, more than half of the students endorsed activities such as listening to English programs, watching English movies, reading English books, and speaking English with native speakers as effective approaches in improving their English skills. More students at School A than at School B believed that reading English books and magazines, listening to English programs, watching English movies, and reading aloud in English were effective approaches (Figure 8). In

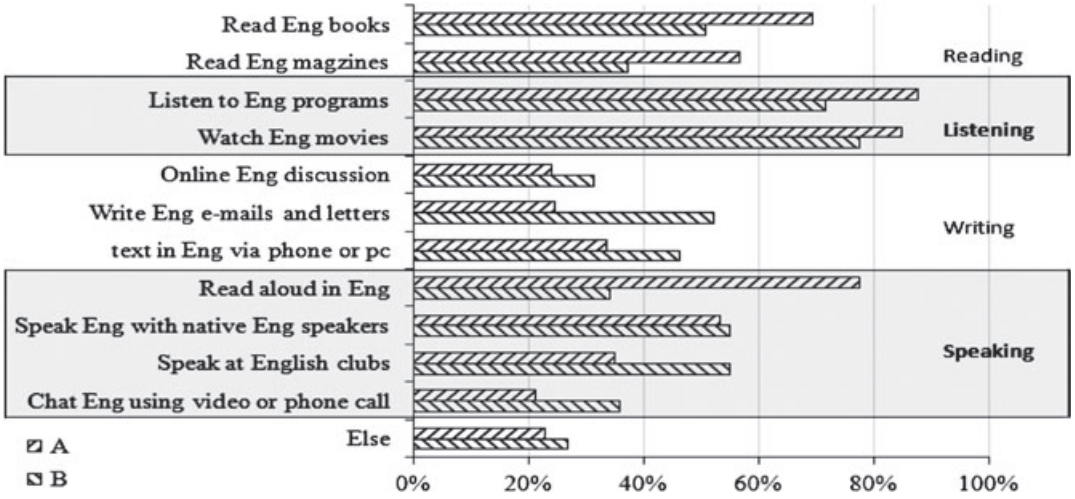

Figure 8 Reported activities that were considered effective for learning English by school.

Table 2 Mean Section Scores on the TOEFL iBT Practice Test by School

School	*N*	Reading (*SD*)	Listening (*SD*)	Speaking (*SD*)	Writing (*SD*)	Total (*SD*)
A	480	13.94 (9.14)	11.50 (8.45)	15.00 (7.68)	16.20 (7.15)	56.64 (27.27)
B	127	6.83 (7.90)	9.86 (7.92)	10.16 (7.49)	7.85 (6.47)	34.42 (24.70)
All	607	12.45 (8.89)	11.16 (8.34)	13.99 (7.64)	14.45 (7.01)	51.99 (26.75)
2011 TOEFL iBT population		20.15 (6.75)	20.05 (6.70)	20.40 (4.60)	21.00 (5.00)	81.50 (20.51)

contrast, more students at School B than School A believed that participating in online English discussions, writing e-mail messages or letters in English, texting in English, speaking in English clubs, and speaking in English on video chats or phone calls were effective (Figure 8).

Results Based on the Longitudinal Sample

Overall, students who participated in this study had relatively low levels of English proficiency, as indicated by their mean section scores on the TOEFL iBT practice tests shown in Table 2. Across all 607 students, the mean scores were 12.45 on the Reading section, 11.16 on Listening, 13.99 on Speaking, 14.45 on Writing, and 51.99 on the total test. All these scores fell at least one standard deviation below the population means for the 2011 TOEFL iBT operational test takers (Educational Testing Service, 2012), which are displayed in the last row of Table 2. The students at School A had higher mean scores on each section than those at school B.

A further analysis based on the students who took both tests found substantial score gains on the section scores and the total scores (Table 3). Across schools, students improved their Reading scores from 8.63 to 16.37 on average ($d = .97$), Listening scores from 8.17 to 13.80 ($d = .77$), Speaking scores from 13.32 to 15.25 ($d = .32$), Writing scores from 13.40 to 16.60 ($d = .53$), and total scores from 39.84 to 56.69 ($d = .57$).

Students at School A improved moderately on the total score ($d = .48$), substantially on the Reading ($d = 1.20$) and Listening sections ($d = .82$), moderately on the Writing section ($d = .47$), but less on the Speaking section ($d = .27$). At School B, students showed substantial improvements on the Reading ($d = .90$), Speaking ($d = 1.05$), and Writing ($d = 1.46$) sections, and moderate on the Listening section ($d = .56$).

To further compare the score gain patterns, the effect sizes of section score changes for schools A and B are depicted in Figure 9, which shows greater score gains on Reading and Listening sections for students at the China-based School A but smaller score gains on the Speaking and Writing sections than those at the US-based School B.

A multivariate repeated GLM was fitted to the data, where the two sets of section scores (the Reading, Listening, Speaking, and Writing scores on the pretest and posttest) were treated as the multivariate outcome variables, and school, gender, weekly number of extracurricular hours reportedly spent on English, and number of reported years learning

Table 3 Section Scores on the Pretest and Posttest by School

School		Reading (*SD*)	Listening (*SD*)	Speaking (*SD*)	Writing (*SD*)	Total *M*(*SD*)
A ($n = 90$)	Pretest	9.82 (7.48)	8.42 (6.51)	13.80 (6.04)	14.61 (5.60)	43.69 (22.22)
	Posttest	18.89 (7.65)	14.45 (8.20)	15.50 (6.38)	17.40 (6.38)	55.88 (27.74)
	d	1.20	0.82	0.27	0.47	0.48
B ($n = 21$)	Pretest	3.78 (4.41)	7.39 (7.00)	1.36[a] (5.65)	7.00 (2.70)	23.74 (19.00)
	Posttest	9.30 (7.31)	11.59 (7.90)	15.10 (3.51)	14.00 (5.99)	46.85 (19.53)
	d	0.90	0.56	1.05	1.46	1.20
All	Pretest	8.63 (7.27)	8.17 (6.49)	13.32 (5.99)	13.40 (5.93)	39.84 (22.68)
	Posttest	16.37 (8.64)	13.80 (8.03)	15.25 (5.93)	16.60 (6.21)	53.69 (26.06)
	d	0.97	0.77	0.32	0.53	0.57

[a]This extremely small number was mainly because more than half of speaking responses had low audio quality and could not be processed through the SpeechRater.

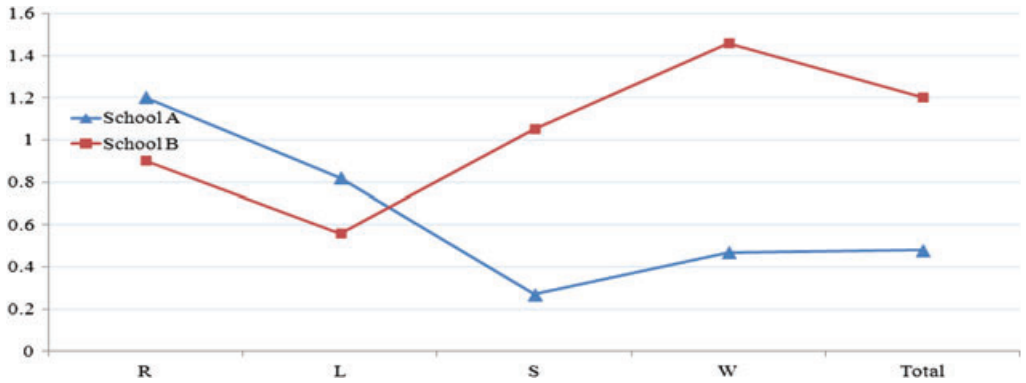

Figure 9 Effect sizes of score improvement by school and TOEFL iBT section. R = Reading; L = Listening; S = Speaking; W = Writing.

English were treated as the predictors. The within-subject main effect associated with the two tests (due to test repeating or learning) was also significant: Wilks' lambda = .29, $F(4, 23) = 14.16$, $p < .001$. The univariate test results suggest that the score increase on each of the four section scores was significant: $F = 27.32$, $p < .001$ for Reading; $F = 18.97$, $p < .001$ for Listening; $F = 6.56$, $p = .017$ for Speaking; $F = 20.90$, $p < .001$ for Writing. Only the multivariate main effect associated with the number of hours spent per week studying English was significant: Wilks' lambda = .30, $F(12, 61) = 2.94$, $p = .003$. Neither the main effects associated with program nor gender was significant.

Figure 10 displays the relationship between score gains and number of extracurricular hours spent on English learning, where greater amounts of time reportedly spent studying outside the classroom each week were associated with higher test scores and greater score improvement in general. However, it also seems that the average score gains on the Speaking and Writing sections were not always the largest for those who reported spending the longest time per week.

To determine the relationship between the English activities students reported engaging in and their English-language proficiency improvement, average gains (in terms of Cohen's *d*) on each section score were computed for each English-learning activity between students who considered it to be effective and those who did not (see Table 4). As the test of program effects was not statistically significant, we analyzed the data of all students across programs together. Students who considered reading English books as an effective way to learn English had greater mean scores on the Reading section ($d = 0.46$) than those who did not believe it to be effective. However, only a small effect size was observed for students who acknowledged the effectiveness of reading English magazines or newspapers as compared to those who did not ($d = .11$; see Table 4). Students who considered watching English movies as an effective way to improve English skills scored higher on the Listening section than those who did not ($d = .59$). A small effect size was observed with regard to listening to English programs ($d = .10$), meaning that students who believed listening to English programs was an effective approach scored only slightly higher on the Listening section than those who did not. Those who considered participating in online written discussions in English, text messaging in English, or writing English e-mail messages or letters to be effective, seemed to have comparable or even lower Writing scores than those who did not, with effect sizes of .00, −.21, and −.04,

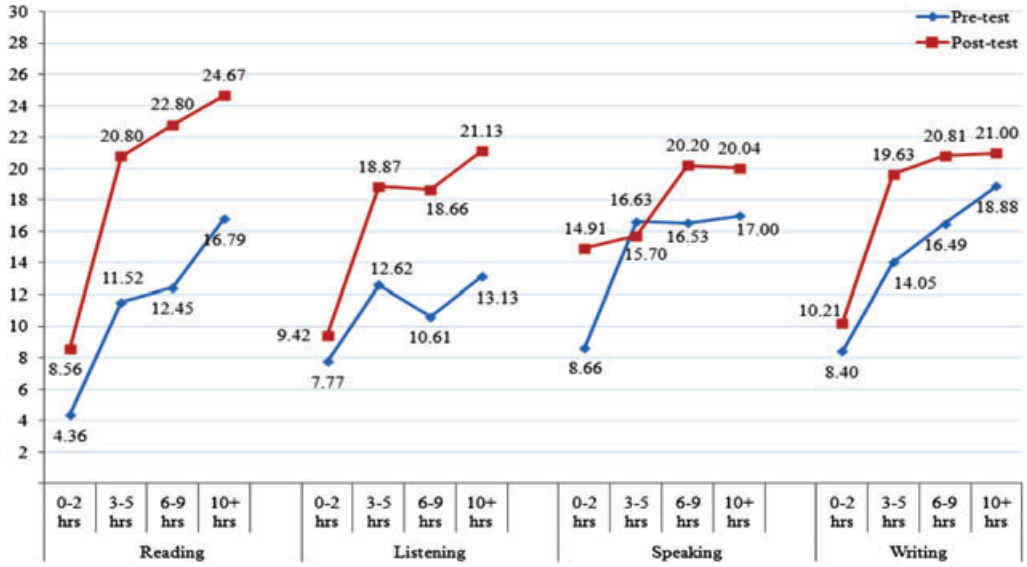

Figure 10 Mean section scores by category of reported hours per week on learning English.

Table 4 Average Section Score Gain Difference by Recognized Effective English-Learning Activity

Effective English-learning activity	*d*[a]	Directly related TOEFL iBT section
Read English books	.46	Reading
Read English magazines	.11	
Listen English programs	.10	Listening
Watch English movies	.59	
Online English discussion	.00	Writing
Write in English emails letters forums	−.21	
Texting in English via phone or computers	−.04	
Read aloud in English	−.37	Speaking
Practice speaking with native English	.32	
Practice speaking at English clubs	.43	
Chat in English via phone or video	−.07	

[a]*d* is the Cohen's effect size used to measure the score gain difference between students who considered the learning activity effective and those who did not.

respectively. Beliefs in the effects of practicing English with native English speakers and at English clubs was positively related to students' Speaking section scores ($d = .32$ and .43, respectively). However, those who believed in the effects of reading aloud in English turned out to have smaller score gain on the Speaking section during the posttest than those who did not believe so ($d = -.37$).

Results Based on the Cross-Sectional Sample

Finally, the data of all students who took the posttest at School A were analyzed. As was mentioned earlier, 116 students were finishing their first-year studies and 119 were finishing their second year at the time of the posttest. As the instructional methods, materials, and student characteristics at School A only changed minimally from 2010 to 2012, when this study was carried out, we believe this cross-sectional data from the school could provide additional evidence to confirm whether learning effects were reflected in the TOEFL iBT practice test scores. In other words, a substantial score difference between the second-year and the first-year students would also support the claim of the TOEFL iBT test being able to reflect the language proficiency improvement over the academic year. The analysis results indicated substantial score differences between the first-year and second-year students, on average. Specifically, the first-year students scored on average 9.97 on the Reading section, much lower than that of the second-year students (18.57, $d = 1.05$); on the Listening section,

Table 5 Multivariate Test Results (*F*-Statistics Values) Based on the Cross-Sectional Sample

Factor	*df*1, *df*2	*F*-statistic	*p*
Grade	4,190	24.22	.001
Gender	4,190	3.96	.004
Grade × Gender	4,190	.136	.969

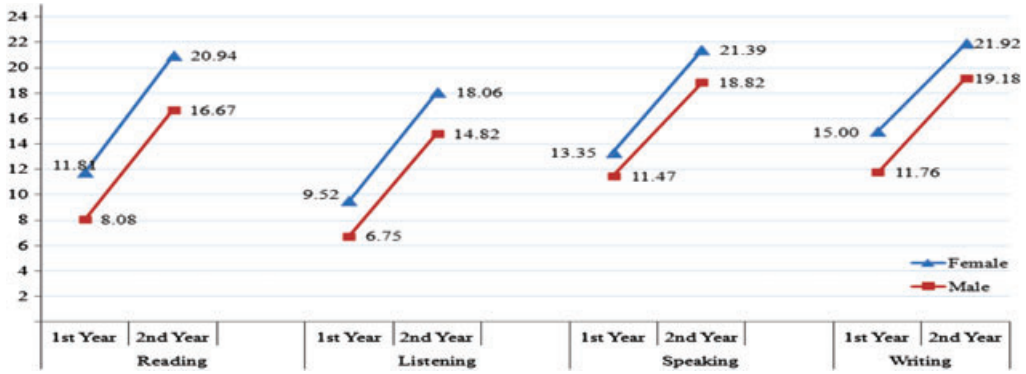

Figure 11 Mean TOEFL iBT section scores by grade and gender based on the cross-sectional sample at School A.

the averages were 8.06 and 16.21, respectively ($d = 1.08$); on the Speaking section, 12.27 and 19.95, respectively ($d = 1.10$); on the Writing section, 13.62 and 20.06, respectively ($d = 1.02$).

A multivariate GLM was fitted to the data, using the four section scores as the outcome variables, and grade level and gender as the predictors. Table 5 provides the results of the analyses. It was found that the multivariate main effects associated with grade (first or second year) and gender were significant (Table 5). However, there was no significant interaction between grade and gender, $F(4, 190) = .136$, $p = .969$.

The main effects related to gender and grade were plotted in Figure 11, where the patterns of score increase between the first-year and second-year students, and the score differences between the two gender groups, were clearly shown on each section.

Further analyses confirmed that the univariate main effects associated with grade and gender were both significant for each of the four section scores. Compared to male students, female students performed moderately better on the Reading ($d = .52$) and Listening ($d = .44$) sections, and slightly better on the Speaking ($d = .27$) and Writing ($d = .23$) sections. The second-year (G2) students performed substantially better on all sections than first-year students (G1), with the effect size >.90 (Table 6).

The number of hours per week that students reportedly spent learning English was entered as a predictor after grade and gender in the multivariate GLM. The multivariate main effect associated with the reported number of hours spent per week on English was not significant, $F(12, 397) = 1.57$, $p = .097$. However, the reported number of hours spent per week on English was positively associated with student scores on the Listening section only, $F = 3.20$, $p = .025$. Finally, the interaction among grade, gender, and reported hours spent per week on learning English was positively associated with student scores on the Reading section, $F = 2.67$, $p = .049$. A further comparison revealed that students who reported spending 0 – 2 hours weekly on English activities earned a lower mean score than students who reported spending 3 – 5 hours weekly,

Table 6 Univariate *F*-Test Results Based on the Cross-Sectional Sample

Main factor	Reading	Listening	Speaking	Writing
Grade	$F = 61.53$ ($p < .001$)	61.83 ($p < .001$)	59.64 ($p < .001$)	69.78 ($p < .001$)
d (G2−G1)	.93	.95	.96	.91
Gender	12.55 ($p < .001$)	8.10 ($p < .005$)	4.96 ($p < .027$)	12.14 ($p < .001$)
d (F−M)	.52	.44	.27	.23

Table 7 Mean (*SD*) of Section Scores by Reported Number of Hours Spent per Week on Learning English

Number of hours	*N*	Reading (*SD*)	Listening (*SD*)	Speaking (*SD*)	Writing (*SD*)
1: 0–2	25	10.02 (12.88)	8.41 (12.20)	14.48 (9.91)	14.90 (8.78)
2: 3–5	68	15.05 (9.46)	12.08 (8.96)	16.66 (7.28)	17.07 (6.45)
3: 6–9	52	15.26 (8.94)	15.74 (8.47)	18.53 (6.88)	18.58 (6.09)
4: 10+	79	16.29 (8.36)	13.40 (7.92)	18.45 (6.43)	18.64 (5.70)
*d*12		0.48	0.37	0.27	0.30
*d*13		0.51	0.75	0.51	0.52
*d*14		0.65	0.55	0.54	0.57
*d*23		0.02	0.42	0.26	0.24
*d*24		0.14	0.16	0.26	0.26
*d*34		0.12	−0.29	−0.01	0.01

on the Reading section ($d = .48$), the Listening section ($d = .37$), and the Speaking section ($d = .27$). Substantial differences were also found between the students who reportedly spent the minimum time on English per week (0–2 hours) and those who reported spending a moderately high amount of time (6–9 hours) per week on the Reading ($d = .51$), Listening ($d = .75$), Speaking ($d = .51$), and Writing ($d = .52$) sections. These differences were slightly smaller than the differences found between the minimum reported weekly time group (0–2 hours) and the maximum reported weekly time group (10 or more hours) on the Reading ($d = .65$), Listening ($d = .55$), Speaking ($d = .54$), and Writing ($d = .57$) sections. The differences between the students who reportedly spent 3–5 hours, 6–9 hours, or more than 10 hours per week were small with regard to most of the section scores. However, there was a moderate difference in the Listening section scores between the students who reported that they spent 3–5 hours per week and those who reported they spent 6–9 hours per week on English-related activities ($d = .42$; Table 7).

To demonstrate the relationship between the reported hours spent per week learning English and the section scores, we plotted the two variables for students grouped by grade in Figure 12. Overall, there was a positive association, such that students who reported spending more hours per week learning English performed better on average than those who reported spending less time, on each section; the positive relationship seemed to be stronger for the Reading and Listening sections than for the Speaking and Writing sections, regardless of grade. These trends were consistent for the first-year students' section scores. For the second-year students, similar trends were observed, except for the Reading and Listening sections. On the Reading section, students who reported they spent 3–5 hours per week learning English outperformed the students who reported they spent more time per week. On the Listening section, students who reportedly spent more than 10 hours per week on English actually scored lower than those who reported spending 6–9 or 3–5 hours per week.

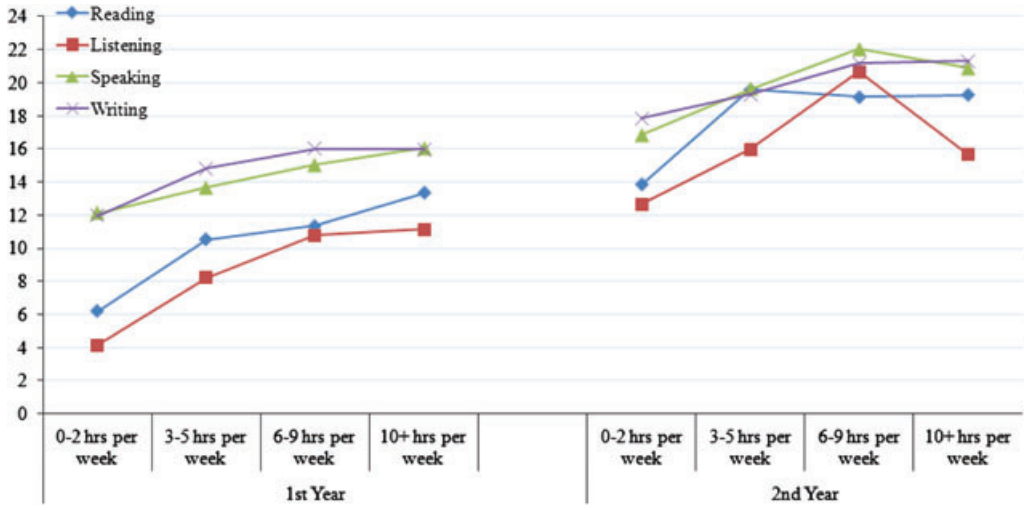

Figure 12 Mean section scores by grade and reported hours per week on learning English.

Discussion

In summary, although we used TOEFL iBT practice tests in inauthentic testing contexts where the results were not used for high-stake decision making (i.e., admission decisions), the test results were nonetheless able to capture students' learning effects as indicated by the test score gains. We found moderate to substantial levels of score gains on the TOEFL iBT practice test for students who were studying either in US-based intensive English programs or in a China-based special purpose high school. This finding was particularly noticeable for the Reading and Listening sections. Students in the US-based English program demonstrated greater improvements on the Speaking and Writing sections compared to those who studied in China. Students who attended the China-based program increased their scores on the Reading and Listening sections more than the students in the US-based programs did.

Additionally, students in China reported greater external pressure (or more motivation) to learn English than those who studied in US-based intensive programs. This was reflected in percentages of students acknowledging learning English because it is required by school or parents and because English is required for admission to further study. Furthermore, students in China reported that they had studied English for a greater amount of time than those already in the United States. Students in the United States and China reportedly spent a comparable amount of extracurricular time reading in English, whereas the latter group reported spending more time practicing listening to English than did the former group. More of the students in the United States indicated they were engaged in English writing- and speaking-related activities than did those in China, possibly accounting for their greater score gains on these sections.

A greater proportion of students in the China-based school than in the US-based English programs believed that reading aloud in English and several receptive skill-related activities, such as reading English books and magazines, listening to English programs, and watching English movies, were effective in improving their English proficiency. On the other hand, a greater proportion of students in the US-based programs than in the school in China believed that participating in communicative skill-related activities, such as online written English discussions, writing English e-mail messages or letters, writing text messages in English, speaking in English clubs, and speaking in English on the phone or video calls, were effective ways to improve their English proficiency.

With regard to the first research question, students who attended these English-learning programs appeared to improve their English-language proficiency as reflected by their TOEFL iBT practice test scores, with moderate to large effect sizes on each of the sections. These students may be motivated to do better by virtue of being in one of the schools that participated in the study. That is, participating in the study may have encouraged the students to try harder and be more willing to take the test a second time. However, as students were unaware of the second test when they were initially tested, this type of motivation-related issue probably had little effect on the current results.

For the second research question, different patterns of score improvement were found among the participating schools. Students in the United States showed greater improvements on the Writing and Speaking sections than those in China, whereas the students in China displayed greater gains on the Reading and Listening sections than those in the United States. However, such differences were found to be not statistically significant.

For the third research question, female students in general outperformed male students on the section scores, but both gender groups achieved comparable levels of improvement. The number of weekly extracurricular hours that students reportedly spent studying English was positively associated with their section scores on the pretest and posttest, as well as the score changes. More specifically, students who reported studying English for 3 or more hours per week had greater mean scores on each section of both test times compared to students who reported studying only 2 or fewer hours per week. Several types of reported English-learning activities had moderate level effects on students' English proficiency levels as reflected by their TOEFL iBT practice scores. Specifically, reading English books was associated with Reading score gains, watching English movies with Listening score gains, and speaking with native English speakers and at English clubs with Speaking score gains.

Several important implications should be noted based on these findings. First, it seems that students can improve their English-language proficiency noticeably as reflected by score gains on the TOEFL iBT practice test within a year, whether they study in a US-based or international intensive English program. Even at programs that did not specifically prepare students for a standardized English test (e.g., TOEFL iBT test), students achieved moderate to large score gains on the practice test, which is encouraging for students who might have previously considered the TOEFL iBT test to be difficult. From this perspective, the study provides empirical evidence in support of the argument that the TOEFL iBT practice test can capture changes (or improvement) in English-language skills as a result of learning and instruction. It also supports

the claim that an individual's English proficiency can be improved at least moderately over the course of several months, as indicated by a standardized test such as the TOEFL iBT test.

Secondly, although the between-program effects were not statistically significant, differential score gains were observed between the US-based English programs and the Chinese program in terms of Cohen's effect size. It seems reasonable to assume that in the immersive context of English programs in the United States, students would attain greater improvements in English proficiency compared to students studying in less immersive contexts, controlling for other factors such as program length and intensiveness. This assumption is supported by the greater US-based student score gains on the Speaking and Writing sections compared to the China-based student gains. However, the finding that the students in China actually achieved greater gains on the Reading and Listening sections suggests that the degree of immersion may play a less important role in affecting students' receptive (reading and listening) abilities than in improving their productive (speaking and writing) abilities. It seems a reasonable argument may be that students in China engage more often in English-learning activities that are not typically (or less frequently) engaged in by English learners in the United States. That is, simply associating with native English-speaking peers (or faculty members who are proficient English users) does not necessarily make students more efficient English learners.

It seems plausible that the program in China may have placed more emphasis on receptive skills in its instruction as well as on students' English-learning activities outside the classroom. In comparison, the US-based programs may have emphasized communicative-oriented English skills such as speaking and writing. Such differences in the instructional and learning foci may be associated with the differential score gain patterns on the four sections of the TOEFL iBT practice test between the programs in the United States and in China, as was found in this study.

The findings of this study also provide some encouraging news in terms of the amount of time that students need to spend learning English in order to improve their English-language skills and obtain a noticeable score gain on the TOEFL iBT test. At School A in China, the students typically needed to finish high school level general education courses in different subjects in their first 2 years. In addition, they took intensive English courses oriented toward TOEFL in the second year. The results based on both the pretest and posttest sample and the cross-sectional sample suggest that, besides the other general educational courses, 9 months of intensive courses can lead to moderate to substantial improvement in English skills and TOEFL iBT practice test scores, especially on the Listening and Reading sections. At School B, although the English courses were mainly for academic purposes and not directly targeted at the TOEFL iBT test or another test, students managed to improve their English-language skills to a moderate degree after 6 months of study. It seems safe to infer that even if their starting level of English proficiency is low, students can improve their English-language skills at least moderately after 6 – 9 months of intensive study.

Finally, the current findings should be evaluated in the context of students' English proficiency level. The students included in this study had low or intermediate English-language proficiency, as indicated by their initial TOEFL iBT practice scores. More advanced students might have shown stronger learning effects than those studied here. This assumption seems consistent with previous findings or arguments (Fan, 2001; Young, 2007). However, it would be interesting to examine in the future whether students starting with different English-language proficiency levels (low to high) improve their skills to the same degree in settings similar to those used in this study.

There are several limitations to be noted for this study. First, we were not able to exert any experimental control over program offerings, and instead relied on existing individual programs' English curricula and methods of instruction. This undoubtedly limits the generalizability of our findings. In addition, the factors related to English proficiency that we chose to study may have been too broad. And the use of students' self-reported extracurricular learning activities may be proved by using a more objective and accurate measure. Future studies might take advantage of the present findings to narrow the scope of the specific English-learning activities or to focus on other program-related factors. For example, the overall time spent weekly on learning English may be narrowed down to the time engaged in English-learning activities related to a specific modality (e.g., reading, listening, speaking, or writing individually or in a combined way). Such information may be helpful to determine whether certain modality-specific English-learning activities are positively associated with score gains in tests that measure those modalities. We believe this might enable a better controlled study that leads to results that are more generalizable. Also, further investigation might be worthwhile to examine the relationship between extracurricular English-learning activities and score change within each program, as the current findings were based on the pooled sample across programs due to the no significant difference between programs on the score gains.

A motivation issue associated with the current sample may exist. As students were invited to take the two practice tests regardless of their reasons for learning English, it is possible that some participants may not have planned to take the operational TOEFL iBT test in the future. These students may have been less motivated to improve their English-language skills, which in turn may have affected their performance on the test. However, as students' motivation is unlikely to have changed dramatically between the pretest and posttest, and the difference between students' scores on the two tests was the primary focus of this study, it seems reasonable to suggest that a lack of student motivation may have affected the study results only to a limited degree.

On the other hand, because of the voluntary nature of participation for this study, students who did take the two tests may have been more strongly motivated to improve their English proficiency and test their learning progress than those who took the practice test only once. This also may have reduced the generalizability of the study results. It is likely that smaller score gains would have been observed had all the students in this study taken both tests.

The constructed responses for the Writing and Speaking sections were scored only using an automated scoring engine, which might affect the scores' accuracy. It might be valuable to compare the results using human raters with those found in this study, which may provide additional information on the topics of this study.

Further, the numbers of both the English-language programs and the participating students were very small in this study compared to the large number of English-language programs and learners around the world. Therefore, one should be cautious in trying to generalize the current findings to programs or student populations different from those of this study.

Nonetheless, despite the limitations discussed above, we believe that the study reported here extends an important aspect of the validity argument for TOEFL iBT scores. The data that we have analyzed show that the TOEFL iBT test is capable of capturing gains in English-language proficiency as a result of formal instruction and reported less formal learning outside the classroom.

Notes

1 e-rater is an automated scoring engine for essays developed at Educational Testing Service (ETS). The e-rater engine is used in combination with human raters to score the Writing sections of the TOEFL and *GRE*® tests, as psychometric research has demonstrated that this combination is superior to either machine scoring or human scoring on their own. See Attali and Burstein (2006) for more details.

2 SpeechRater is an automated scoring engine of spoken responses developed at ETS. It is used to score spontaneous responses, in which the range of valid responses is open-ended rather than narrowly determined by the item stimulus. SpeechRater has been used to score speaking responses to the TOEFL Practice Online test since 2006. See Xi, Higgins, Zechner, and Williamson (2011) and Zechner and Xi (2008) for details.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning and Assessment*, *4*, 1–29.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1–26). New York, NY: Routledge.

Chin-Chin, K. (2007). EFL listening comprehension strategies used by students at the Southern Taiwan University of Technology. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *68*(3A), 978.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Educational Testing Service. (2012). *Test and score data summary for TOEFL iBT tests and TOEFL PBT tests*. Princeton, NJ: Author.

Eggly, S., & Schubiner, H. (1991, April). *A course in speaking fluency for foreign medical residents in the United States*. Paper presented at the Eastern Michigan University Conference on Languages and Communication for World Business and the Professions, Ypsilanti, MI.

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. *International English Language Testing System Research Reports*, *4*, 207–254.

Ellis, R. (1993). *The study of second language acquisition*. England: Oxford University Press.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). England: Oxford University Press.

Fan, Y. (2001). *Plateau of EFL learning*: *A psycholinguistic and pedagogical study*. Retrieved from http://wlkc.nbu.edu.cn/jpkc_nbu/daxueyingyu/download/014.pdf

Forster, D. E., & Karn, R. (1998, March). *Teaching TOEIC/TOEFL test taking strategies*. Paper presented at the meeting of the Teachers of English to Speakers of Other Languages, Seattle, WA.

Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policy & Practice*, *14*(1), 75 – 97.

Jiang, X. (2007). The impact of graphic organizer instruction on English-as-a-foreign-language college students' reading comprehension. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, *68*(3A), 979.

Ling, G., & Bridgeman, B. (2013). Writing essays on a laptop or a desktop computer: Does it matter? *International Journal of Testing*, *13*(2), 105 – 122.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5 – 11.

Mori, S. (1999). The role of motivation in the amount of reading. *Temple University Japan Working Papers in Applied Linguistics*, *14*, 51 – 68.

Park, G. P. (1995). Language learning strategies and beliefs about language learning of university students learning English in Korea. *Dissertation Abstracts International*, *56*(06), 2102A(UMI No.9534918).

Rea-Dickins, P., & Scott, C. (2007). Washback from language tests on teaching, learning, and policy: Evidence from diverse settings. *Assessment in Education: Principles, Policy & Practice*, *14*, 1 – 7.

Ross, S. J. (2005). The impact of assessment method on foreign language proficiency growth. *Applied Linguistics*, *26*(3), 317 – 342.

Tanaka, K., & Ellis, R. (2003). Study-abroad, language proficiency, and learner beliefs about language learning. *Japan Association for Language Teaching Journal*, *25*(1), 63 – 85.

Wilson, K. M. (1987). *Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language* (Research Report No. RR-87-03). Princeton, NJ: Educational Testing Service.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2011). A comparison of two scoring methods for an automated speech scoring system. *Language Testing*, *29*(3), 371 – 394.

Young, S. (2007, December). Effects of instructional hours and intensity of instruction on NRS level gain in listening and speaking. *Center for Applied Linguistics* (*CAL*) *Digest*. Retrieved from http://www.cal.org/resources/digest/digest_pdfs/levelgain.pdf

Zechner, K., & Xi, X. (2008). Towards automatic scoring of a test of spoken language with heterogeneous task types. In J. Tetreault, J. Burstein & R. De Felice (Eds.), *Proceedings of the third workshop on innovative use of NLP for building educational applications* (pp. 98 – 106). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology//W/W08/W08-0912.pdf

Action Editor: James Carlson

Reviewers: Cathy Wendler and Lawrence Davis

E-RATER, ETS, the ETS logo, GRE, LISTENING. LEARNING. LEADING., TOEFL, TOEFL IBT, and TOEIC are registered trademarks of Educational Testing Service (ETS). SPEECHRATER is a trademark of ETS. All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/