Listening. Learning. Leading.®

Research Report
ETS RR–14-15

Investigating the Relationship Between Test Preparation and *TOEFL iBT*® Performance

Ou Lydia Liu

December 2014

ETS Research Report Series

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

Investigating the Relationship Between Test Preparation and *TOEFL iBT*® Performance

Ou Lydia Liu

Educational Testing Service, Princeton, NJ

This study investigates the relationship between test preparation and test performance on the *TOEFL iBT*® exam. Information on background variables and test preparation strategies was gathered from 14,593 respondents in China through an online survey. A Chinese standardized English test was used as a control for prior English ability. Multiple regression analyses were used to investigate the relationship of coaching school attendance and test preparation strategies with TOEFL iBT total scores. Coaching school attendance had little or no relationship with TOEFL test scores across language domains. Confirmatory factor analyses revealed that general English learning strategies and test-specific strategies represent two distinct factors of test preparation. Implications of the findings for test developers and test sponsors are discussed.

Keywords Coaching; language testing; test preparation; TOEFL iBT

doi:10.1002/ets2.12016

Test preparation for high-stakes tests has received extensive attention from test developers, test sponsors, and score users. Among the studies investigating the types of test preparation and/or the effect of preparation strategies on test scores, most have been focused on college and graduate school admissions tests, such as the *SAT*® exam (e.g., Becker, 1990; Briggs, 2004; Kulik, Bangert-Drowns, & Kulik, 1984; Messick & Jungeblut, 1981; Powers, 1987, 1993; Powers & Rock, 1999; Stricker, 1982), ACT (Moss, 1995; Scholes & Lain, 1997), and the *GRE*® General Test (e.g., Powers & Swinton, 1982, 1984; Swinton & Powers, 1983). However, very little is known about the strategies examinees use to prepare for the *TOEFL iBT*® exam, one of the most widely used tests required for the admission of international students to academic programs in English-speaking countries. Furthermore, insufficient research evidence exists documenting the relationship between the strategies used by TOEFL iBT test takers and their TOEFL iBT scores. As the *TOEFL*® exam has established itself as an instrumental tool in global language testing among more than 8,000 institutions and millions of test takers, it is necessary to understand how test preparation predicts test performance and affects the interpretation of TOEFL iBT scores.

Literature Review

Test preparation is defined as "any intervention procedure specifically undertaken to improve test scores, whether by improving the skills measured by the test or by improving the skills for taking the test, or both" (Messick, 1982, p. 70). This definition captures the quintessential purpose of any test preparation activity: that is, to improve test scores. Cole (1982) summarized six possible components of test preparation including being supplied with correct answers (cheating), taking practice tests, maximizing motivation, coping with test anxiety, increasing test wiseness, and instructing test content. These components represent a wide range of activities that test takers can pursue during their test preparation and capture both cognitive and noncognitive aspects of test preparation.

Anastasi (1981) further grouped test preparation into three broad categories based on the nature of the preparation instruction: (a) test-taking orientation, which helps examinees become familiar with the testing procedures and overcome anxieties due to the strangeness of the test; (b) coaching, which usually involves intense and short-term practice on similar item formats offered by either commercial companies or by school-based programs; and (c) training in broadly applicable cognitive skills, which contributes to overall improvement of cognitive ability and thus to enhanced test performance. Investigators of the effect of test preparation on test performance largely focus on the first and second categories, as the effect of the third category, although most desirable and valid, is difficult to measure.

Corresponding author: O. L. Liu, E-mail: lliu@ets.org

Several meta-analytic studies have investigated the effect of test preparation. Becker (1990) synthesized findings from 48 studies evaluating the effect of SAT coaching and concluded that the average effect of coaching is about 9 points on SAT verbal scores and 19 points on SAT math scores on a 200–800 scale. Kulik, Kulik, & Bangert (1984) expanded the meta-analysis to include studies of all available aptitude tests that have a control group. Based on results from 35 papers containing 38 studies, Kulik, Bangert-Drowns, et al. (1984) and Kulik, Kulik, et al. (1984) concluded that coaching programs in general have a positive impact on test performance. However, the effect of coaching is significantly smaller on the SAT (with a mean effect size of .15) than on other testing programs (with a mean effect size of .43). Data used in the Kulik et al. study were reanalyzed in a few subsequent studies (Kulik & Kulik, 1986; Pearlman, 1984; Witt, 1992). Pearlman (1984) challenged the results of the Kulik, Bangert-Drowns, et al. (1984) and Kulik, Kulik, et al. (1984) study by suggesting that the observed difference in coaching effect between SAT and non-SAT tests was merely the result of sampling error. However, he did find differences when looking at year of publication. Studies published prior to 1940 had a larger effect size than studies published from 1952 on. Kulik and Kulik (1986), on the other hand, questioned the accuracy of the meta-analysis techniques that were used in the Pearlman (1984) study. They argued that if the right formulas derived from Hedges and Olkin (1985) would be used, the sampling error would only account for 12% of the variance on the SAT instead of 51%, as claimed by Pearlman (1984). Cole (1982) reviewed five experimental studies involving a control group related to the effect of test preparation on SAT scores and found that the average growth effect is estimated to be 26 points for SAT Verbal and 21 points for SAT Math on the 200–800 SAT score scale.

Besides the above meta-analysis studies, individual studies have also yielded mixed findings about the effect of coaching on high-stakes tests. Powers and Rock (1999) analyzed data from 4,200 SAT test takers, of whom 507 had attended either commercial coaching programs or programs offered by schools, using six statistical models to triangulate the results. The estimated coaching effect ranged from 4 to 14 points on SAT Verbal and from 12 to 22 on SAT Math on a 200–800 scale across the six methods. Powers and Rock concluded that there is an effect of coaching, but the effect is far less prominent than that boasted by commercial coaching companies. Briggs (2004) conducted a causal investigation of commercial coaching and SAT performance and found that depending on the covariates included in the analysis (e.g., age, socioeconomic status, parent education), the estimated coaching effect could range from none to about 69 points on the SAT Verbal and from 30 to 79 points on the SAT Math, both on a 200–800 score scale. Scholes and Lain (1997) investigated the effect of three test prep strategies on ACT performance: taking a practice test, using workbooks, and taking a test preparation course. Results from 69,251 first-time test takers on the ACT showed that taking a practice test had a positive but small effect on the ACT composite score (a .40 increase in score, effect size less than .10). However, the two other strategies showed a negative impact on composite score (−.60 for both) after controlling for high school GPA and grade level. Scholes and Lain (1997) concluded that engaging in test preparation activities will not yield large gains in ACT scores.

Compared to the general consensus from research on admissions tests that there is a small effect of coaching on test scores, the effect of coaching on language tests is less clear. Most of the studies investigating the effect of test preparation on language tests have focused on two major testing programs, the TOEFL and the International English Language Testing System (IELTS). Nguyen (2007) investigated the effect of a preparation course on both the TOEFL iBT Listening and the IELTS Listening tests. The IELTS preparation course ran 10 weeks with about 1.5–2 daily hours devoted to listening training. The TOEFL preparation course ran 2 weeks with 4 hours daily of training at one school and 4 weeks with 2 daily hours training at another school. The IELTS group performed significantly better on the IELTS Listening test than the TOEFL group, but the two groups performed equally on the TOEFL test. Nguyen concluded that the effect of test preparation is more obvious on the IELTS than on the TOEFL iBT. Other studies also reported a significantly positive relationship between test preparation and scores on IELTS Listening (e.g., Hayes & Read, 2004), Writing (e.g., Brown, 1998), and Speaking (e.g., Issitt, 2008) tests.

Green (2007) investigated the impact of three types of preparation courses on the IELTS Writing scores: (a) courses specifically targeting the IELTS, (b) courses targeting general academic writing, and (c) courses combining the two. With a pre- and posttest design, the author found that students in all three courses significantly improved their IELTS scores ($p < .01$), but students in the first and third types of courses did not make more score gains than students in the second type of writing courses.

Bachman, Davidson, Ryan, and Choi (1995) investigated the effect of a preparation course for the First Certificate in English (FCE) test on both FCE and TOEFL. They conducted multiple regressions using either FCE or TOEFL score as the dependent variable, and test preparation and the corresponding TOEFL or FCE score as the covariate. They found that a preparation course on FCE accounted for 6–9% of the variance in FCE scores but had very little impact on TOEFL scores.

Ward and Xu (1994) investigated the effect of instruction on summarization skills on TOEFL scores. After a 6-week training on summarization skills with written materials, students had a .5 standard deviation score gain on TOEFL scores. The authors also found that students who used summarization skills in an English as a second language class had a significantly larger score gain on the TOEFL test than students who did not use summarization skills.

Swain, Huang, Barkaoui, Brooks, and Lapkin (2009) examined the strategies students use in taking the speaking section of the TOEFL iBT and found that the relationship between strategy and test performance varied significantly by task and by type of strategy (i.e., affective, metacognitive, communicative). Certain metacognitive strategies such as self-correcting were significant predictors of TOEFL iBT speaking scores.

Purpose of This Study

This study aims to investigate how TOEFL iBT test takers' preparation strategies are associated with their test scores. The literature review identified three main limitations with many existing studies investigating the effect of test preparation on language tests: (a) the sample sizes of the studies tend to be small, as often only students who are language learners are included in the studies; (b) the observations are often limited to one or two institutions or programs without much generalizability; and (c) the investigations are often focused on only one skill area without providing an overall picture of how test preparation may influence test scores across the four language domains (i.e., reading, writing, listening, speaking). To overcome these limitations, this study included a large sample of TOEFL iBT test takers from hundreds of institutions and investigated all four language skills.

Methods

Selection of Study Participants

Ideally, any investigation of the relationship between strategy use and test scores should include all TOEFL iBT test takers. However, a potential problem is that there is no standardized control for prior English proficiency for all test takers. Because the use of strategies is most likely a self-selected behavior, high-proficiency examinees may use strategies more effectively than low-proficiency examinees. Purpura (1998) found that high- and low-proficiency language learners may use the same strategies in taking a reading test, but the strategies may have a differential impact on their test performance. Therefore, a control for examinees' prior English proficiency is needed to eliminate the confounding between language proficiency and the effectiveness of strategies. In mainland China, a standardized English test, called the College English Test Band 4 (CET4), is required for all college students. Chinese test takers represent about 20% of the TOEFL iBT population, and more than 50% of them are either college or graduate students who have likely taken the CET4. Given the large percentage of Chinese examinees among the TOEFL population and the well-known intense efforts that Chinese examinees devote to test preparation, it is reasonable to assume that the strategies used by Chinese test takers are similar to, if not more comprehensive than, the strategies used by test takers in other countries.

Survey Construction

To gather information on TOEFL iBT test preparation strategies, a survey was designed and sent to test takers who had recently completed the test. In gathering test preparation strategies, we reviewed strategies covered in popular test preparation materials offered by coaching schools (e.g., the New Oriental School) for the TOEFL iBT. We also reviewed literature for commonly used test preparation strategies for English language tests (Brown, 1998; Elder & O'Loughlin, 2003; Issitt, 2008). We focused on test preparation strategies that are frequently practiced, meaningful (as opposed to guessing or taking advantage of test construction flaws [Cole, 1982]), and legitimate (e.g., not cheating). Two types of strategies emerged from the review: (a) general practice strategies that aim to improve test takers' overall English ability as well as their performance on the TOEFL iBT (Anastasi, 1981; Cole, 1982; Messick, 1981); and (b) content-based preparation specifically targeting the test, which usually involves intense and short-term practice on similar item formats. Practicing with the *Step-by-Step English* tapes to improve overall listening proficiency is an example of the first category, and practicing with the TOEFL Practice Online (*TPO*™ practice tests) is an example of the second category.

The survey consisted of demographic information, TOEFL iBT-related background information (e.g., whether the examinee is a first-time test taker), general preparation strategies for improving English language ability (e.g., reading English magazines), and test-specific preparation strategies (e.g., using the TPO). Respondents were asked to indicate the frequency of their practice using a 5-point Likert scale (i.e., *seldom*, *a couple of times a year*, *a couple of times a month*, *a couple of times a week*, and *almost every day*).

The survey also asked the test takers to report their score on the CET4. As mentioned earlier, the CET4 is an English language test required for all college students in China. It measures English language skills in listening, reading, and writing. Many students also report CET4 scores for graduate school or job applications, as sufficient English skills are considered essential for college graduates in China. Since test takers may have taken the CET4 after they took TOEFL and those CET4 scores would not be a proper control for TOEFL scores, the survey specifically asked the test takers to report their CET4 scores before they had taken the TOEFL.

An interview was conducted with 15 TOEFL iBT examinees to further understand the strategies they had used in preparing for the TOEFL iBT. The interviewees commented on the clarity, coverage, and format of the survey questions. The final survey was created online with 52 questions and took about 15 minutes to complete. In addition, the online survey was designed to avoid missing data.

Data Collection

The data used in this study consist of survey data and test score data. The survey data were collected from TOEFL iBT test takers, and the official TOEFL iBT test scores were obtained from the ETS data warehouse. The survey and test data were linked using the unique e-mail addresses that test takers provided when they took the test.

To collect the survey data, e-mails were sent to Chinese test takers who (a) took the TOEFL iBT no earlier than 6 months before the data collection (March 2010), which took place between September 1, 2009, and February 28, 2010, (b) provided a valid e-mail address, and (c) gave consent to ETS to contact them for research purposes. Based on these three criteria, e-mails were sent to 90,488 test takers. The e-mails directed the test takers to the online survey. It was made clear that responses to the survey would not affect the test takers' current or future TOEFL iBT scores and that the survey was only for research purposes. To attract participation from test takers and to ensure timely data collection, the e-mail specified that a monetary incentive equivalent to $50 in the form of a gift card would be provided to the first 500 respondents. The online survey was shut off after about 4 weeks. Responses were collected from 14,593 test takers in mainland China.

Analyses

Descriptive analyses were conducted for student background variables (e.g., gender), TOEFL iBT background variables (e.g., test repeater status, purpose for taking the test), general English learning strategies, and test-specific preparation strategies. Analysis was also conducted to determine whether the respondents' test performance was representative of the performance of all TOEFL iBT test takers.

Confirmatory factor analysis (CFA) was conducted to see whether the two kinds of preparation strategies (general vs. test specific) represented two distinct factors of test preparation. CFA was performed in LISREL 8.8 (Joreskog & Sorbom, 1993) using weighted least squares estimation with polychoric correlations and asymptotic covariance matrices as input. Several model fit indices were used to evaluate the fit of the CFA models, including the comparative fit index (CFI), non-normed fit index (NNFI), and root mean square error of approximation (RMSEA; Muthen & Muthen, 2004). Reliability (Cronbach's alpha) of each strategy scale was calculated. The correlation between the two kinds of strategies was also provided. Results from the two-factor CFA were compared to those of a one-factor CFA to determine the suitability of the two-factor model.

Stepwise regression analyses were conducted to investigate the relationship between coaching school attendance, course taking, and TOEFL iBT performance. The survey asked about six types of courses commonly offered by coaching schools: vocabulary, reading, listening, speaking, writing, and overall test simulation courses.[1] Owing to the high multicollinearity among the coaching school attendance and course-taking variables (i.e., one has to attend the coaching school to enroll in the courses; it's likely for someone to enroll in multiple courses), separate regression analyses were conducted for the coaching school attendance and the six course-taking variables. Specifically, in the first set of regression models, coaching school attendance and CET4 scores were used as the predictors. The regression analysis was repeated five times with total score and the four skill scores as the dependent variables, respectively. In the second set of regression models, the six course-taking variables and CET4 scores were used as the predictors. Similarly, the regression analysis was repeated five times with total score and the four skill scores as the dependent variables, respectively. CET4 scores were included in all of the regression models to control for prior English language proficiency. Note that since 12% of the sample was high school students and only 1% had taken the CET4, the students without CET4 scores were excluded from the regression analyses.

To further understand the relationship between specific test preparation practice and TOEFL iBT test performance, separate stepwise regression analysis was conducted, with the TOEFL iBT total score and the four skill scores as the outcomes variables, respectively, and all the general and specific strategies as predictors. All the strategy variables were included as predictors, as they are frequently used by TOEFL iBT test takers with the premise that these strategies will help to improve test scores. CET4 scores were included as a control for English language proficiency in all the regression analyses. Any missing data were treated with pairwise deletion.

Results

Descriptive Statistics

The 14,593 respondents included 47% males. See Table 1 for the descriptive information of the test takers. The respondents were outperformed by the TOEFL population in terms of the TOEFL total score ($p<.01$). However, the significance was probably due to the large sample size. The effect size was .04, suggesting no practical difference. Comparison between the respondents and the 83,294 nonrespondents on the TOEFL total score also showed a significant difference favoring the respondents, but again the effect size was small ($d=.09$).

Of the respondents, 76% ($n=11{,}030$) had taken the CET4 test required for college students in China by the time they took the TOEFL. The CET4 was modified in 2005. The score scales for the old test had a mean of 72 and a standard deviation of 12. The score scales for the new test have a mean of 500 and a standard deviation of 70. Among the 11,030 CET4 test takers, 20% took the old test and 80% took the new test. The old scale scores were transformed to the new scale scores using the score concordance table provided by the Chinese Ministry of Education.

About 49% of the respondents (see Table 1) reported that they had attended coaching schools when preparing for the TOEFL iBT, including 49.4% males. Coaching school attendants included 29.5% of the high school students, 44.7% of the college students, 72.8% of the graduate students, and 70.7% of the working professionals. See Figure 1 for courses taken at the coaching schools. More than 80% took each of the reading, writing, listening, and speaking courses.

Use of General Strategies

Among the general test preparation strategies that students used, listening to English programs (e.g., radio) and watching movies in English were the two most frequently used strategies (Table 2). Of the respondents, 31% reported that they listened to English programs twice a week and 25% reported that they practiced English this way almost every day. Forty-two percent reported that they watched movies in English twice a week, and 25% reported that they did so almost every day. The least used practice strategies included practicing spoken English with a native speaker and practicing English at English salons, probably because both strategies require the presence of other people and, therefore, are not as convenient as practicing by oneself.

Use of Test-Specific Strategies

Some of the most frequently used TOEFL iBT-specific test preparation strategies included taking notes while listening (87%), identifying main ideas while listening (85%), identifying topic sentences, main points, and key statements in reading (82%), skipping unfamiliar words in reading (82%), and improving fluency in speaking (82%; Table 3). The least practiced strategy was taking notes while reading (40%). About 47% of the respondents indicated that they used the TPO offered by ETS to prepare for the TOEFL iBT.

Table 1 Descriptive Statistics by Test-Taker Characteristics

			Total TOEFL score	
	n	%	Mean	*SD*
Gender				
Males	6,816	47	70.7	27.0
Females	7,768	53	71.5	27.1
Number of times taking the test				
First time	8,802	60	70.1	27.1
Twice	2,789	20	73.1	27.5
Three times	1,753	12	73.1	26.6
Four times	808	5	70.9	26.5
More than four times	432	3	68.7	25.1
Academic status				
High school	1,748	12	67.8	27.9
College	8,644	59	72.0	27.2
Graduate school	2,976	20	70.8	25.8
Professional	1,138	8	71.0	27.8
Other	71	1	71.0	28.1
Reason for taking the test				
Admission to high school	552	4	61.6	26.8
Admission to college	2,736	19	68.2	27.4
Admission to graduate school	10,009	69	72.9	26.8
Improve English proficiency	489	3	67.2	26.9
Career advancement	634	4	67.2	26.4
Other	173	1	68.1	26.3
Time spent on preparation				
1–2 weeks	2,798	19	74.0	28.0
3–4 weeks	3,935	27	72.5	27.0
5–6 weeks	2,498	17	70.5	27.1
7–8 weeks	2,294	16	70.4	26.5
More than 8 weeks	3,059	21	67.7	26.4
Coaching school attendance				
Yes	7,195	49	72.7	27.4
No	7,389	51	69.5	26.7

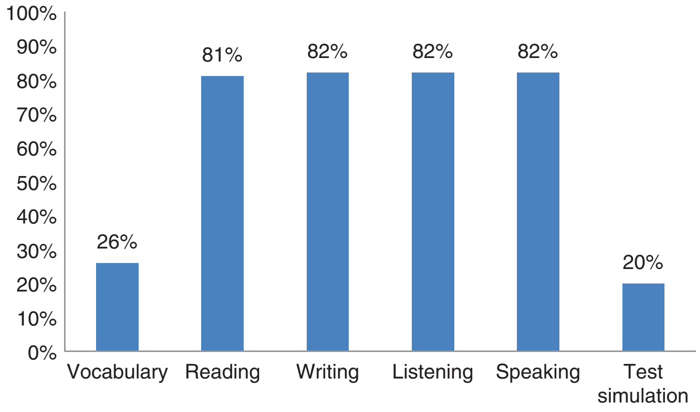

Figure 1 Course taking at the coaching school.

Factor Analysis

CFA was conducted to investigate whether the general and specific strategies represented two distinct factors (Table 4). Results were also compared to a one-factor solution. Model fit was evaluated using the criteria specified by Muthen and Muthen (2004). For acceptable fit: CFI/NNFI > .90, RMSEA < .08; for good fit: CFI/NNFI > .95, RMSEA < .06. The two-factor model showed better fit than the one-factor model: (CFI = .97 vs. .83; NNFI = .98 vs. .85; RMSEA = .04 vs. .09).

Table 2 Use of General English Learning Strategies

	Seldom		A couple of times a year		A couple of times a month		A couple of times a week		Almost every day	
Item text	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Read English books	2,615	18	2,714	19	3,019	21	3,547	24	2,689	18
Read English magazines	3,038	21	2,892	20	4,366	30	3,085	21	1,203	8
Listen to English programs (e.g., radio)	1,261	9	1,621	11	3,470	24	4,557	31	3,674	25
Watch movies in English	636	4	1,006	7	3,364	23	6,133	42	3,444	24
Participate in online discussions in English (e.g., online forum, online text chatting)	6,146	42	2,825	19	3,162	22	1,731	12	719	5
Write e-mails, letters, and diaries in English	2,438	17	2,676	18	3,917	27	3,442	24	2,110	14
Read English aloud to improve spoken English	2,671	18	2,358	16	3,921	27	3,578	25	2,055	14
Practice spoken English with a native English speaker	6,161	42	3,132	21	2,598	18	1,781	12	910	6
Practice spoken English at English salons	8,213	56	2,666	18	2,029	14	1,221	8	453	3

Table 3 Use of TOEFL iBT-Specific Test Preparation Strategies

	Yes		No	
Item text	*n*	%	*n*	%
Skim reading (i.e., read the text quickly to obtain a general impression of what the text is about)	11,680	80	2,904	20
Identify topic sentences, main points, and key statements in reading	11,990	82	2,594	18
Skip unfamiliar words in reading	11,903	82	2,681	18
Take notes while reading	5,763	40	8,821	60
When practicing reading items, look at the questions first and then go back to the reading passage	10,274	70	4,310	30
Identify main ideas while listening	12,446	85	2,138	15
Pay attention to transitional phrases in listening	10,954	75	3,630	25
Take notes while listening	12,622	87	1,962	13
Listen to topics/themes similar to the TOEFL Listening section	11,405	78	3,179	22
Improve pronunciation and intonation in speaking	10,748	74	3,836	26
Improve fluency in speaking	11,938	82	2,646	18
Express opinions in an organized structure in spoken English	11,049	76	3,535	24
Practice spoken English on topics similar to the TOEFL Speaking section	11,297	77	3,287	23
Practice spoken English using templates (e.g., use common transitional phrases; use common argument structure)	9,690	66	4,894	34
Write on topics similar to the TOEFL Writing test	10,665	73	3,919	27
Practice writing using templates (e.g., use common transitional phrases, use common expressions)	9,579	66	5,005	34
Memorize vocabulary for the TOEFL test	10,841	74	3,743	26
Practice using TOEFL simulation test or released TOEFL items	11,458	79	3,135	21
Practice using the TPO	6,821	47	7,771	53

For the two-factor structure, the Cronbach's α was .78 for the nine questions on general strategies, and .81 for the 19 questions on TOEFL iBT-specific strategies. These two factors showed a moderate correlation ($r = .31$).

The Relationship Between Coaching School Attendance, Courses Taken, and Test Performance

Table 5 provides both standardized and unstandardized regression coefficients in addition to the *R*-squared values from the stepwise regression analyses. A standardized coefficient stands for the expected change in the dependent variable per one standard deviation increase in the predictor. An unstandardized coefficient stands for the change in the independent variable per one unit change in the predictor. In the case of this study, both are provided, as readers may be interested in knowing both the relative size of the relationship (i.e., the standardized coefficients) and the change in score points per one unit change in the binary variables (e.g., whether or not one had attended a coaching school).

Only significant predictors from the stepwise regression are included in Table 5. After controlling for CET4 score, coaching school attendance was a significant predictor of both the total TOEFL iBT score and the four subscale scores.

Table 4 Factor Loadings of the Two-Factor Confirmatory Factor Analysis Model

	Mean	*SD*	F1	F2
Read English books	3.07	1.37	0.60	
Read English magazines	2.76	1.23	0.67	
Listen to English programs (e.g., radio)	3.53	1.22	0.62	
Watch movies in English	3.74	1.03	0.41	
Participate in online discussions in English (online forum/text chatting)	2.18	1.23	0.68	
Write e-mails, letters, and diaries in English	3.01	1.29	0.60	
Read English aloud to improve spoken English	3.00	1.31	0.66	
Practice spoken English with a native English speaker	2.19	1.27	0.70	
Practice spoken English at English salons	1.84	1.14	0.65	
Skim reading (read the text quickly to obtain a general impression of text)	1.20	0.40		0.57
Identify topic sentences, main points, and key statements in reading	1.18	0.38		0.62
Skip unfamiliar words in reading	1.18	0.39		0.42
Take notes while reading	1.60	0.49		0.37
When practicing reading items, look at the questions first and then go back to the reading passage	1.30	0.46		0.33
Identify main ideas while listening	1.15	0.35		0.57
Pay attention to transitional phrases in listening	1.25	0.43		0.54
Take notes while listening	1.13	0.34		0.54
Listen to topics/themes similar to the TOEFL Listening section	1.22	0.41		0.60
Improve pronunciation and intonation in speaking	1.26	0.44		0.61
Improve fluency in speaking	1.18	0.39		0.63
Express opinions in an organized manner in spoken English	1.24	0.43		0.62
Practice spoken English on topics similar to the TOEFL Speaking section	1.23	0.42		0.66
Practice spoken English using templates (e.g., use common transitional phrases; use common argument structure)	1.34	0.47		0.65
Write on topics similar to the TOEFL Writing test	1.27	0.44		0.63
Practice writing using templates (e.g., use common transitional phrases, using common expressions)	1.34	0.47		0.64
Memorize vocabulary for the TOEFL test	1.26	0.44		0.45
Practice using TOEFL simulation test or released TOEFL items	1.21	0.41		0.57
Practice using the TPO	1.53	0.50		0.64

The unstandardized coefficients show that coaching school attendance was associated with an increase of 1.86 score points for total score (on a 0–120 scale), which is fairly small. The strongest relationship to coaching was observed with reading, with 1.01 points increase (on a 0–30 scale). Furthermore, compared to the standardized coefficients of CET4 scores (i.e., ranging from .30 to .50), the standardized coefficients for coaching school attendance were also fairly small, ranging from .03 to .09, which suggests that attending coaching school had a relatively small effect on TOEFL iBT performance.

The relationship between course taking and test performance showed similar patterns. After controlling for CET4 scores, taking a vocabulary course significantly predicted both total and subscale scores, as did taking a test simulation course, except in the case of speaking. Taking a reading course significantly predicted reading scores, and taking a listening course and a speaking course both significantly predicted listening scores. Speaking appeared to be the most difficult domain to improve through taking training courses. The effect of course taking appeared to be fairly small in general. Only two unstandardized coefficients were larger than 1 score point out of 30 total points for the subscales: a simulation course on reading and a listening course on listening. The standardized regression coefficients were also small (i.e., ranging from .03 to .07).

The Relationship Between Strategy Use and Test Performance

The results from five separate stepwise multiple regression analyses in Table 6 show the association between individual strategies and TOEFL iBT performance. Given the large number of predictors included in the multiple regressions, multicollinearity was examined using the variance inflation factor (VIF; Longnecker & Ott, 2004). A common criterion is that a VIF less than 2.5 indicates no or negligible multicollinearity. The VIF values met this criterion for the five regression analyses, ranging from 1.03 to 1.76 (mean = 1.30 and *SD* = .19).

Table 5 Results From the Multiple Regression Models

	Total score			Writing			Reading			Listening			Speaking		
	B	β	R^2	**B**	β	R^2	**B**	β	R^2	**B**	β	R^2	**B**	β	R^2
CET4[a]	.11***	.30	.078	.03***	.50	.206	.03***	.43	.190	.04***	.45	.206	.02***	.41	.167
Coaching school attendance	1.86***	.03	.079	.40***	.03	.208	1.01***	1.00	.200	.70***	.04	.208	.16**	.04	.167
CET4[b]	.11***	.30	.078	.03***	.45	.206	.03***	.43	.100	.04***	.44	.206	.02***	.41	.167
Vocabulary course	3.60***	.03	.080	.70***	.06	.210	.86***	.05	.197	.69***	.03	.208	.45***	.05	.168
Writing course															
Reading course							.62***	.06	.200						
Listening course										1.06***	.05	.209			
Speaking course										.73***	.07	.210			
Simulation course	2.02*	.04	.081	.49***	.03	.211	1.01***	.06	.202	.76***	.06	.211			

Note. **B** = unstandardized regression coefficient; β = standardized regression coefficient; R^2 = R-squared values suggesting the proportion of variance in the outcomes variable explained by the current predictor and all previous predictors.
[a]Coaching school attendance and CET4 scores were the predictors. The regression analysis was repeated for total score and subscale scores. [b]The six course-taking variables and CET4 scores were the predictors. The regression analysis was repeated for total score and subscale scores. *$p < .05$. **$p < .01$. ***$p < .001$.

Table 6 presents the unstandardized regression coefficients from the multiple regression models. Only significant coefficients are included. The standardized coefficients of the five regression models ranged from .01 to .09, indicating small effect sizes. The R-squared value of the regression analyses was .09 for total scoring, .25 for reading, .27 for writing and listening, and .24 for speaking.

As expected, CET4 scores were the strongest predictor of TOEFL performance for both total and scale scores. In terms of general English learning strategies, writing e-mails, letters, and diaries in English and practicing spoken English at English salons were shown to be positive predictors of both total score and skill scores, respectively. Reading English books and reading English magazines positively predicted total and subscale scores except for speaking. Participating in online discussions predicted reading scores. Both listening to English programs and watching movies in English were associated with higher scores in listening. Both reading English aloud and practicing spoken English with a native speaker were associated with improved scores in reading, listening, and speaking.

In examining the relationship of test-specific strategies with TOEFL iBT performance, a main finding was that the relationship tended to be domain specific. For example, the strategies targeting reading (e.g., take notes while reading) were most likely to correlate positively with reading performance. Strategies targeting listening (e.g., paying attention to transitional phrases in listening) were most likely to predict listening scores. Similarly, writing strategies were positively associated only with writing scores (e.g., writing on topics similar to those used in the TOEFL Writing test).

An exception was that strategies aiming to improve speaking scores appeared to predict both speaking and listening scores. For example, improving pronunciation and intonation in speaking and improving fluency in speaking were associated with improved scores in both speaking and listening. Memorizing vocabulary for the TOEFL test, practicing using a TOEFL simulation test or released TOEFL items, and using the TPO were strong predictors of both total and skill scores, respectively.

Discussion

A limitation of this study is that the CET4 scores used were self-reported scores rather than those obtained from official sources. The 14,593 respondents came from hundreds of institutions over 30 provinces in China. Obtaining permission to access test takers' official CET4 scores from the institutions would have been extremely difficult, if not impossible.

A second limitation lies in the lack of information on test takers' coaching school attendance for English tests other than the TOEFL iBT. Although about half of the respondents reported not attending coaching schools for the TOEFL iBT, it was unclear whether they attended coaching schools for other English tests (e.g., GRE, the *TOEIC*® test, CET4) or for

Table 6 Unstandardized Parameter Estimates from the Multiple Regressions

	Total score	Writing score	Reading score	Listening score	Speaking score
CET4 score	0.12***	0.04***	0.03***	0.04***	0.02***
Read English books	1.71*	0.67*	0.73***	0.42*	
Read English magazines	1.85*	0.73*	0.81*	0.54***	
Listen to English programs (e.g., radio)				0.86***	
Watch movies in English				0.63**	
Participate in online discussions in English (online forum/text chatting)			1.04*		
Write emails, letters, and diaries in English	2.84***	1.34***	0.79***	0.57***	0.20***
Read English out aloud to improve spoken English			0.62***	0.47***	0.66***
Practice spoken English with a native English speaker			0.54**	0.78**	0.79***
Practice spoken English at English salons	1.98***	0.89**	0.41***	0.25***	0.73***
Skim reading (read the text quickly to obtain a general impression of text)	1.87**		1.21**		
Identify topic sentences, main points, and key statements in reading			0.82***		
Skip unfamiliar words in reading			0.63*		
Take notes while reading	1.71*	0.72***	0.60***	0.44***	
When practicing reading items, look at the questions first and then go back to the reading passage			0.04***		
Identify main ideas while listening				0.64***	
Pay attention to transitional phrases in listening				0.56**	
Take notes while listening	1.81*			1.03***	
Listen to topics/themes similar to the TOEFL Listening section	1.49*			1.21***	
Improve pronunciation and intonation in speaking				0.78*	0.83**
Improve fluency in speaking	1.71***			0.69**	1.34***
Express opinions in an organized manner in spoken English					
Practice spoken English using templates (e.g., use common transitional phrases, use common argument structure)	2.11**				1.52**
Write on topics similar to the TOEFL writing test		1.71***			
Practice writing using templates (e.g., use common transitional phrases, use common expressions)		0.84**			
Memorize vocabulary for the TOEFL test	2.90***	1.29***	1.06***	0.40**	0.24***
Practice using TOEFL simulation tests or released TOEFL items	2.98***	0.91***	1.16***	0.62***	0.23***
Use TOEFL Practice Online (TPO)	1.85***	0.62**	0.53*	0.33*	0.26**

Note. Unstandardized parameter estimate from the multiple regression models.
$^{*}p < .05.$ $^{**}p < .01.$ $^{***}p < .001.$

other purposes (e.g., English training for career advancement). Coaching for other tests or for other purposes may also have had an impact on TOEFL iBT scores, and this issue awaits future research.

A third limitation is that no item-level data were available for this study, and therefore we were not able to provide the reliability and intercorrelations of the four skill areas. However, readers are referred to several TOEFL iBT validity studies for such information (e.g., Attali, 2011, on writing; Liu, 2011, on reading; Sawaki & Nissan, 2009, on listening; Sawaki, Stricker, & Oranje, 2009, on all four skills). According to the Sawaki et al. (2009) study, the correlations among the reading, listening, and writing sections ranged from .86 to .89. The interfactor correlations between the speaking factor and the other three factors ranged from .66 to .82. It is important to note that no causal inference is implied in the findings, as there is no randomization in this study. All of the reported relationships are correlational.

Despite these perceived limitations, we believe that this study provides useful information on the relationship between test preparation and TOEFL performance. A major strength of this study is its use of scores from the operational administrations of the TOEFL. Many previous studies used published test forms or TOEFL-like tests to investigate the effect of coaching. Students may not have exerted their best effort in these testing situations. A recent meta-analysis study shows that motivation could have an impact as large as .64 *SD* on test performance (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). Motivation would not be an issue for this study, as the TOEFL is a high-stakes test for most test takers. A second strength of the study is its inclusion of CET4 as a control for prior English proficiency. Having a standardized control variable of such a large scale has not been commonly seen in previous studies. In addition, the CET4 is also a high-stakes test, so students' test motivation would not be questionable. In this study, CET4 has shown to be a strong and consistent predictor of both total and scale scores for TOEFL iBT scores.

Two clear findings emerge from the investigations in this study: First, coaching has little or no relationship with TOEFL iBT scores, depending on the language domain; second, the general and test-specific strategies represent two distinct types of test preparation behaviors.

After controlling for the test takers' scores on the CET4, attending coaching school was a significant predictor of both total and subscale scores. However, attending coaching school was only associated with an increase of 1.86 (**B**, unstandardized coefficient) score points in the TOEFL total score, on a 0–120 scale. As for the four subscales, even the largest effect showed a small increase of 1.01 (**B**) points on reading on a 0–30 scale. Coaching has nearly no effect on speaking (**B** = .16) and negligible effect on writing (**B** = .40) and listening (**B** = .70).

The relationship of course taking with test performance is also weak. Vocabulary and test simulation courses appear to be the most effective courses in predicting TOEFL performance. Some of the strongest effects include taking vocabulary course on total score (**B** = 3.60), taking a course on listening (**B** = 1.06), and taking a test simulation course on total score (**B** = 2.02) and on reading (**B** = 1.01). All other effects are fairly small without any practical significance.

In conclusion, coaching has a fairly weak relationship with the reading and listening skills assessed by the TOEFL iBT and has almost no relationship with writing and speaking. The finding is consistent with previous studies on test preparation concerning the TOEFL (e.g., Bachman et al., 1995; Nguyen, 2007) and other standardized tests (e.g., Powers & Rock, 1999).

The most significant difference between the general and test-specific strategies is whether the preparation focuses only on the TOEFL iBT or on general English learning. Compared to the often short-term and intense test-specific preparation strategies, general English learning strategies tend to be less intense and may take a longer time to show any effect on test scores. Test-specific strategies may bring about immediate effects on test scores, but the effect may be limited to one specific skill domain. For instance, in this study, practicing writing on topics similar to the TOEFL Writing test was associated only with and increased only writing scores. On the contrary, general English learning strategies, if practiced consistently and frequently, may be significantly correlated with improved English abilities in multiple skill domains. For example, writing e-mails, letters, and diaries in English positively predicted both total scores and the four skill scores.

Findings on the relationship between various strategies and test scores (Table 6) have implications for test developers and test sponsors. First, practicing with TOEFL-like tests seems to be effective in predicting test scores (e.g., listening to topics similar to those on the TOEFL Listening section, writing on topics similar to those on the TOEFL Writing test). In order to reduce the likelihood of inflated scores gained by excessive drills with TOEFL-like tests, test developers may consider increasing the variety of topics when designing reading, writing, listening, and speaking tests. Introducing new topics from time to time may help measure examinees' real ability when they have not seen such topics before. Second, when defining and refining the test constructs, test developers may want to consider the relationship between test scores and test takers' English ability in contexts other than the TOEFL iBT. For example, should a test taker with a high score on the TOEFL Writing be expected to write clear and effective e-mails regarding academic issues in English or to be able to articulate his/her opinions in an academic online discussion? These considerations may benefit test developers when they make decisions about construct levels, assessment formats, and scoring rules. Third, as practicing using TOEFL simulation tests or released TOEFL items has the largest effect on TOEFL total score (**B** = 2.98), the test sponsor may want to inform test takers of that strategy in the official test guide and increase access to such preparation materials for the purpose of providing equal opportunities to all test takers.

Note

[1] In the overall test simulation courses, instructors provide training experiences to test takers, which take use of TOEFL iBT simulation tests and focus on both test content and test-taking strategies (e.g., test-taking pace, omission strategies).

References

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist*, *36*(10), 1086–1093.

Attali, Y. (2011). *Automated subscores for TOEFL iBT® independent essays* (Research Report No. RR-11-39). Princeton, NJ: Educational Testing Service.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An investigation of comparability of two tests of English as a foreign language*. Cambridge, England: Cambridge University Press.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research*, *60*, 373 – 417.

Briggs, D. C. (2004). Causal inference and the Heckman model. *Journal of Educational and Behavioral Statistics*, *29*(4), 397 – 420.

Brown, J. D. H. (1998). Does IELTS preparation work? An application of the context-adaptive model of language program evaluation. *IELTS Research Reports*, *1*, 20 – 37.

Cole, N. (1982). The implications of coaching for ability testing. In A. Wigdor & W. Garner (Eds.), *Ability testing: Uses, consequences and controversies* (pp. 389 – 414). Washington, DC: National Academies Press.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(19), 7716 – 7720.

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

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

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. W. Cheng, Y. J. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 97 – 111). Mahwah, NJ: Erlbaum.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Issitt, S. (2008). Improving scores on the IELTS speaking test. *ELT Journal*, *62*(2), 131 – 137.

Joreskog, K. G., & Sorbom, D. (1993). *LISREL8: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ: Erlbaum.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, *95*, 179 – 188.

Kulik, J. A., & Kulik C. L. C. (1986). *Operative and interpretable effect sizes in meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. Retrieved from ERIC database. (ED 276759)

Kulik, J. A., Kulik, C. L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*, 435 – 447.

Longnecker, M. T., & Ott, R. L. (2004). *A first course in statistical methods*. Belmont, CA: Thomson Brooks/Cole.

Liu, O. L. (2011). Does major field of study and cultural familiarity affect TOEFL® iBT reading performance? A confirmatory approach to differential item functioning. *Applied Measurement in Education*, *24*(3), 235 – 255.

Messick, S. (1981). The controversy over coaching: Issues of effectiveness and equity. In B. F. Green (Ed.), *Issues in testing: Coaching, disclosure, and ethnic bias* (pp. 21 – 53). San Francisco, CA: Jossey-Bass.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist*, *17*, 67 – 91.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin*, *89*, 191 – 216.

Moss, G. (1995). *The effects of coaching on the ACT scores of African-American students*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. Retrieved from ERIC database. (ED 399265)

Muthen, L. K., & Muthen, B. O. (2004). *Mplus: The comprehensive modeling program for applied researchers. User's guide* (3rd ed.). Los Angeles, CA: Muthen & Muthen.

Nguyen, T. N. H. (2007, May). *Effects of test preparation on test performance – the case of the IELTS and TOEFL iBT listening tests*. Paper presented at Teaching English to Speakers of Other Language (TESOL) in the Internationalization of Higher Education in Vietnam, Hanoi, Vietnam.

Pearlman, K. (1984). Validity generalization: Methodological and substantive implications for meta-analytic research. In H. Wing (Chair), *Meta-analysis: Procedures, practices, and pitfalls*. Symposium conducted at the meeting of the American Psychological Association, Toronto, Ontario, Canada.

Powers, D. E. (1987). Who benefits the most from preparing for a "coachable" admissions test? *Journal of Educational Measurement*, *24*(3), 247 – 262.

Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice*, *12*, 24 – 30.

Powers, D., & Rock, D. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement*, *36*(2), 93 – 118.

Powers, D. E., & Swinton, S. S. (1982). *The effects of self-study of test familiarization materials for the analytical section of the GRE Aptitude Test* (Research Report No. RR-82-17). Princeton, NJ: Educational Testing Service.

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology*, *76*(2), 266 – 278.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test-takers: A structural equation modeling approach. *Language Testing*, *15*, 333–379.

Sawaki, Y., & Nissan, S. (2009). *Criterion related validity of the TOEFL iBT listening section* (TOEFL iBT Research Report 08). Princeton, NJ: Educational Testing Service.

Sawaki, Y., Stricker, L., & Oranje, A. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*(1), 5–30.

Scholes, R. J., & Lain, M. M. (1997, March). *The effects of test preparation activities on ACT Assessment scores*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. Retrieved from ERIC database. (ED 409341)

Stricker, L. J. (1982). *Test disclosure and retest performance on the Scholastic Aptitude Test* (College Board Report No. 82-7). Princeton, NJ: Educational Testing Service.

Swain, M., Huang, L., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT (SSTiBT): Test-takers' reported strategic behaviors* (TOEFL iBT Research Report No. 10). Princeton, NJ: Educational Testing Service.

Swinton, S. S., & Powers, D. E. (1983). A study of the effects of special preparation on GRE Analytical scores and item types. *Journal of Educational Psychology*, *75*(1), 104–115.

Ward, A. M., & Xu, L. (1994). *The relationship between summarization skills and TOEFL scores*. Paper presented at the annual meeting of the Teachers of English to Speakers of Other Languages, Baltimore, MD. Retrieved from ERIC database. (ED 394332)

Witt, E. A. (1992). *Meta-analysis and the effects of coaching for aptitude tests.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA. Retrieved from ERIC database. (ED 358120)

Suggested citation:

Liu, O. L. (2014). *Investigating the relationship between test preparation and TOEFL iBT® performance* (Research Report No. RR-14-15). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12016

Action Editor: Daniel Eignor

Reviewers: Brent Bridgeman and Donald Powers