Measuring the Power of Learning.™

Research Memorandum
ETS RM–15-09

Examining Performance Differences on Tests of Academic English Proficiency Used for High-Stakes Versus Practice Purposes

Lin Gu

Xiaoming Xi

September 2015

ETS Research Memorandum 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 Memorandum 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 Memorandum 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.

Examining Performance Differences on Tests of Academic English Proficiency Used for High-Stakes Versus Practice Purposes

Lin Gu and Xiaoming Xi
Educational Testing Service, Princeton, New Jersey

September 2015

Corresponding author: L. Gu, E-mail: LGU001@ets.org

Suggested citation: Gu, L., & Xi, X. (2015). *Examining performance differences on tests of academic English proficiency used for high-stakes versus practice purposes* (Research Memorandum No. RM-15-09). Princeton, NJ: Educational Testing Service.

Action Editor: James Carlson

Reviewers: Ikkyu Choi and Spiros Papageorgiou

Abstract

The *TOEFL iBT*® test, developed by Educational Testing Service (ETS), is a high-stakes test widely used for admission to institutions of higher education in English speaking countries. A practice test, the *TOEFL Practice Online TPO*™ test, is the only official practice test for the TOEFL iBT test, and it uses retired TOEFL iBT test forms. Both tests are Internet-based. The two tests are essentially the same, with the TOEFL iBT test administered in a secure, proctored environment and the TPO test given in an unproctored environment. However, they differ significantly in test use. The TOEFL iBT test is commonly used for making high-stakes admission decisions, and the TPO test is a low-stakes practice test. One of the potential uses of TPO scores is to help test takers gauge their future performance on the TOEFL iBT test, although substantiating such a claim of test use still awaits empirical support. The goal of this study is to investigate the relationship between test performance on the TPO test, a low-stakes practice test, and test performance on the subsequent high-stakes TOEFL iBT test. The results of the study can be used to evaluate whether TPO results provide a reasonable prediction of future *TOEFL*® performance, and to advise score users on how to appropriately interpret practice test scores in preparation for the TOEFL iBT test.

Key words: TPO, TOEFL iBT, expectancy-value theory, high-stakes, low-stakes, test consequence

One significant aspect that distinguishes the *TOEFL Practice Online TPO*™ test from the *TOEFL iBT*® test is the impact the scores have on the user, referred to as *test consequence*. Understanding the impact of test consequence on test performance has long been a focus of study in educational research. Studies in this domain are usually framed within the expectancy-value theory (Eccles, 1993; Pintrich, 1988, 1989; Pintrich & De Groot, 1990; Wigfield, 1994), which is a general framework for conceptualizing student motivation within educational settings. Broadly, the theory posits that an individual's motivation to perform, in general, a task is determined by the likelihood of successful task completion as well as the perceived value received in return for succeeding (Wigfield, 1994). Pintrich's expectancy-value model (Pintrich, 1988, 1989; Pintrich & De Groot, 1990), an adaptation of the general framework, proposed three motivational components. The expectancy component concerns a student's prediction of the likelihood of success based upon self-estimation of ability. The value component pertains to a student's perceived importance of successful task completion. The affective component refers to a student's affective and emotional reactions to the task. When applying this model in a testing situation, Wolf and Smith (1995) observed that the value component was largely determined by a test taker's perceived test consequence, suggesting that in a high-stakes testing situation, test takers are predicted to perceive greater importance of test results and, consequently, become more motivated to succeed. Conversely, when the results of a test are considered to be less consequential, test takers are predicted be less likely to put forward the same motivated effort. In essence, the expectancy-value theory projects that students will perform better on high-stakes tests than on tests with lower test consequence because test takers are anticipated to be more motivated and therefore more willing to put forward their best efforts in high-stakes rather than low-stakes testing situations.

Prior studies provided empirical support for this prediction by comparing test performance under conditions of varying consequences. A meta-analysis conducted by Wise and DeMars (2005) examined and found across 12 empirical studies that students, when approaching tests that bear greater consequences, performed better than those who took tests of lesser consequence. The mean difference in each study was examined in standard deviation units and evaluated based on effect size. Cohen (1992) suggested that standardized mean differences of 0.20, 0.50, and 0.80 indicate small, medium, and large effect size, respectively. Wise and DeMars found an average standardized mean difference of 0.59, indicating a medium effect size.

The results from the meta-analysis resonated with the findings from other researchers. Based on the performance on a child development course, Wolf and Smith (1995) found a significant effect for test consequence, with a modest effect size of 0.26. A significant performance difference with a large effect size of 1.27 was found by Napoli and Raymond (2004) between graded and nongraded testing conditions on a college psychology course. Using data from a general education exam under low- and high-stakes test conditions, Cole and Osterlind (2008) detected a small, but significant effect of test consequence. In their study, after controlling for gender and prior academic achievement, the greatest performance difference was found in mathematics and the smallest performance difference was found in science. By experimentally manipulating conditions of test consequence, Liu, Bridgeman, and Alder (2012) found that the students in the two treatment conditions, in which a student outcomes test had either personal or institutional consequence, performed significantly better than those in the control condition, in which the test results bore little consequence.

By broadening the expectancy component in Pintrich's model, Wolf, Smith, and Birnbaum (1995) proposed an expanded version of the model by arguing that the perceived amount of effort needed for success might differentially moderate the relationship between test performance and test consequence. According to this perspective, the expectancy component in Pintrich's model is largely an index of item difficulty in relation to self-estimated ability level. The perceived amount of effort for completing an item, however, relates to how mentally taxing the item is perceived to be by a test taker. Wolf et al. argued that the level of mental taxation is theoretically distinct from the degree of item difficulty. They suggested a model in which the expectancy component consisted of two elements: the likelihood of success if an attempt is made at a problem (item difficulty) and the amount of effort needed to arrive at a correct response (mental taxation). This model predicted that the level of mental taxation moderates the relationship between test performance and test consequence.

Items of varying degrees of mental taxation are commonly classified based on response format (DeMars, 2000; Liu et al., 2012; Sundre & Kitsantas, 2004). Generally, selected-response items are considered to be less mentally taxing, whereas constructed-response items are seen to demand more mental effort from test takers. Response format, selected versus constructed, were found to differentially affect student performance under consequential and nonconsequential conditions in Sundre and Kitsantas and in Liu et al. Wolf et al. (1995) showed that scores on

items deemed more mentally taxing were affected more by test consequence. DeMars also found an interaction between item format and test consequence. She found that students performed better under the high-stakes testing condition than under the low-stakes testing condition, but performance on the constructed-response items exhibited a larger difference than performance on the multiple-choice items. The results of these studies demonstrated the differential impact of response format on the relationship between test performance and test consequence.

In sum, previous studies have provided evidence of the impact of test consequence on test performance. In the context of this study, the TPO test was a low-stakes test and the TOEFL iBT test was a high-stakes test. By taking into account the impact of test consequence on test performance, we examined how TPO test performance is related to TOEFL iBT scores and what factors may play a role in moderating the relationships between the two tests.

Research Objectives

We first investigated the relationship of test performance under high-stakes and low-stakes conditions, taking into account item response format. The expectancy-value theory and its expanded version provided a theoretical basis on which performance under conditions with different consequences and on items with different response formats can be compared. The low-stakes TPO test and the high-stakes TOEFL iBT test both contain selected-response items and constructed-response items. We anticipated that test consequence and item response format could impact how performance on the TPO test was related to scores on the TOEFL iBT test.

Two additional factors, the time interval between test administrations and the mode of testing, were also examined in relation to performance differences across the two tests. We hypothesized that time interval could play a role in moderating the relationship between the tests. As the interval between the practice and the actual test increased, so would the likelihood of language learning or attrition. We also anticipated that the differences in TPO testing mode could affect the relationship between the two tests. Unlike the timed TOEFL iBT test, the TPO test can be taken using either a timed or an untimed testing mode as test takers prefer. The timed mode simulates the TOEFL iBT test-taking experience, using the same timing restrictions as those of the TOEFL iBT test. The untimed mode provides a low-stress environment for users to experience the test, allowing them to proceed at their own pace. In this mode, test takers are not constrained by any time limits to complete the test or parts of the test.

To achieve these goals, we examined scores from the same group of test takers on the TPO test and on the subsequent TOEFL iBT test, and we advanced the following two research questions (RQs):

- RQ1: To what extent do test performances of the low-stakes TPO test and the high-stakes TOEFL iBT test relate to each other on items of different response formats?
- RQ2: To what extent does the time interval between the two tests and the TPO testing mode impact the relationship between scores on the two tests?

Method

Tests

The TOEFL iBT test is a high-stakes test widely used for admission to institutions of higher education in English speaking countries. More information about the test can be found in the TOEFL iBT Research Insight Series (Educational Testing Service, n.d.). The TPO test uses retired TOEFL iBT test forms. Both tests have four sections: reading, listening, speaking, and writing. Reading and listening items have a selected-response format, and speaking and writing items have a constructed-response format. For both tests, scores are reported for each of the four sections in addition to an overall composite score. The score scale is 0−120 for the total test, and 0−30 for each section. For the TOEFL iBT test, the speaking section is scored by human raters, and the writing section by one human rater and the *e-rater*® scoring engine serving as the second rater. The writing and speaking sections of the TPO test are scored by e-rater and *SpeechRater*SM respectively (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012; Xi, Higgins, Zechner, & Williamson, 2008). The different ways of scoring the constructed-response items have implications for the study, which will be discussed later.

Data Source

The study participants were test takers who completed one form of the TPO test, and who subsequently took the TOEFL iBT test between June 2010 and January 2012. A total of 2,342 test takers were included in the data set. The average age of these participants was 27. They were evenly distributed across gender and were from a total of 135 countries. The seven most frequently spoken native languages in order of the number of its speakers were Spanish, German, Chinese, Portuguese, French, Japanese, and Arabic. Native speakers of these seven languages

made up more than half (53.4%) of the total sample. The data for each test taker consisted of (a) scaled reading, listening, speaking, and writing scores from both tests, (b) test-taking dates of both tests, and (c) test-taking mode of the TPO test.

Three groups were identified based on the interval length between the two test administrations. Among the 2,342 test takers, 1,755 took the TOEFL iBT test within 15 days after taking the TPO test (short-interval group); 180 participants took the TOEFL iBT test within 30 days after taking the TPO test (medium-interval group); and the remaining 407 people took the TOEFL iBT test longer than 30 days after taking the TPO test (long-interval group). The test takers could also be grouped by TPO testing mode: 2,105 test takers took the TPO test using the timed mode and 237 test takers took the test using the untimed mode. Table 1 shows the number of subjects by interval and mode.[1]

Table 1. Numbers of Participants by Interval and Mode

Mode	Short	Medium	Long	Total
Timed	1,587	154	364	2,105
Untimed	168	26	43	237
Total	1,755	180	407	2,342

Analyses

Descriptive analyses were conducted. The mean change score, the effect size of the mean change score, and the correlation between TPO and *TOEFL*® scores were calculated. Change scores were inspected in light of standard errors of difference.[2] The distributions of change scores were also examined.

To answer RQ1, a repeated-measures ANOVA was then performed to examine the impact of test consequence on the relationship between the two tests. To respond to RQ2, a general linear model (GLM) analysis was conducted to examine the extent to which the hypothesized factors, time interval and the TPO testing mode, affected the relationship between the two tests. In the GLM analysis, the within-subject variable was testing condition with two levels: consequential and nonconsequential. The between-subject variables were interval and mode. The analysis was performed on both between-subject variables simultaneously.

All analyses were performed based on each of the section scores and the total score separately. The purpose of conducting a separate analysis for each outcome measure was to permit an examination of the impact of response format on the relationship between the tests.

Results

Descriptive Analysis

Table 2 presents the summary statistics for the TPO and TOEFL iBT test performance. Means and mean change scores[3] are shown first, followed by standard deviation. The next column displays the effective sizes of mean change scores. Effect size is a scale-free measure and therefore can be used to compare the relative size of the mean differences across different sections of the test. Effect size is calculated by using pooled variances. Pearson correlations of the section and the total scores between the two tests are shown in the last column.

Table 2. Summary Statistics

N = 2342	Mean	SD	ES	Corr
Reading TPO	22.04	6.43		
Reading TOEFL	23.34	5.65		
Reading change score	1.30	4.56	0.21	0.72
Listening TPO	22.39	6.52		
Listening TOEFL	24.15	5.05		
Listening change score	1.76	4.77	0.30	0.69
Speaking TPO	21.08	2.98		
Speaking TOEFL	23.38	3.75		
Speaking change score	2.30	3.53	0.68	0.47
Writing TPO	22.23	5.48		
Writing TOEFL	23.16	4.15		
Writing change score	0.93	4.24	0.19	0.64
Total TPO	87.74	16.98		
Total TOEFL	94.03	15.97		
Total change score	6.29	9.92	0.38	0.82

Note. Corr = Pearson correlations, ES = effect size.

Mean change scores across all measures were positive, meaning that, on average, test-takers scored higher on the TOEFL iBT test than on the TPO test. The effective size of score changes ranged from 0.19 to 0.68. Speaking performance had the largest standardized score

difference between the two tests. The average score change on the writing section, however, had the smallest effect size. The total score correlation was 0.82, and the correlations for the four sections ranged from 0.47 to 0.72. The correlation between TPO and TOEFL iBT speaking scores was the weakest among all correlations, indicating that speaking performance was the least consistent across the tests. This low correlation could be, in part, explained by the different methods used for scoring speaking between the two tests, which will be discussed later.

Table 3 presents the number and percentage of the subjects whose change scores were within one and two standard errors of difference. For example, on reading, 66% of the changes was within 1 standard error of difference (*SED)*, and 89% within 2 *SED*.

Table 3. Standard Errors of Difference

Section	SEM	1 SED	*N* change within 1 *SED*	Percent change within 1 *SED*	2 SED	*N* change within 2 *SED*	Percent change within 2 *SED*
Reading	2.37	3.35	1,539	66%	6.70	2,042	87%
Listening	2.30	3.25	1,499	64%	6.50	2,004	86%
Speaking	1.58	2.23	1,013	43%	4.46	1,673	71%
Writing	2.37	3.35	1,641	61%	6.70	2,091	89%
Total	4.36	6.17	1,131	48%	12.34	1,789	76%

Note. N = 2,342. SED = standard error of difference, SEM = standard error of measurement.

As shown in Table 3, the percentages of change scores within 2 *SED* was high, close to 90%, for reading, listening, and writing. In contrast, only 71% of the speaking change scores were within 2 *SED*.

Table 3. Standard Errors of Difference

Section	SEM	1 SED	*N* change within 1 *SED*	Percent change within 1 *SED*	2 SED	*N* change within 2 *SED*	Percent change within 2 *SED*
Reading	2.37	3.35	1,539	66%	6.70	2,042	87%
Listening	2.30	3.25	1,499	64%	6.50	2,004	86%
Speaking	1.58	2.23	1,013	43%	4.46	1,673	71%
Writing	2.37	3.35	1,641	61%	6.70	2,091	89%
Total	4.36	6.17	1,131	48%	12.34	1,789	76%

Note. N = 2,342. SED = standard error of difference, SEM = standard error of measurement.

The majority (74%) of the participants (*N* = 1,741) took the timed TPO test and subsequently took the TOEFL iBT test within 30 days. Mean score difference analysis in relation to standard errors of difference was conducted for this group. As shown in Table 4, in general the

percentages of score changes within one and two standard errors of difference were slightly higher than those based on the whole sample, and the general pattern across outcome measures was the same as the one based on the whole sample.

Table 4. Standard Errors of Difference for a Subgroup of Participants

Section	*SEM*	1 *SED*	*N* change within 1 *SED*	Percent change within 1 *SED*	2 *SED*	*N* change within 2 *SED*	Percent change within 2 *SED*
Reading	2.37	3.35	1,203	69%	6.70	1,568	90%
Listening	2.30	3.25	1,185	68%	6.50	1,538	88%
Speaking	1.58	2.23	723	42%	4.46	1,219	70%
Writing	2.37	3.35	1,266	73%	6.70	1,581	91%
Total	4.36	6.17	899	52%	12.34	1,389	80%

Note. N = 1,741. SED = standard error of difference, SEM = standard error of measurement.

The distributions of change scores for the four sections and for the total test are shown in Figures 1 through 5. The horizontal scale reflects the standard errors of difference. As shown in these figures, all distributions were largely symmetrical, approximating a bell shape.

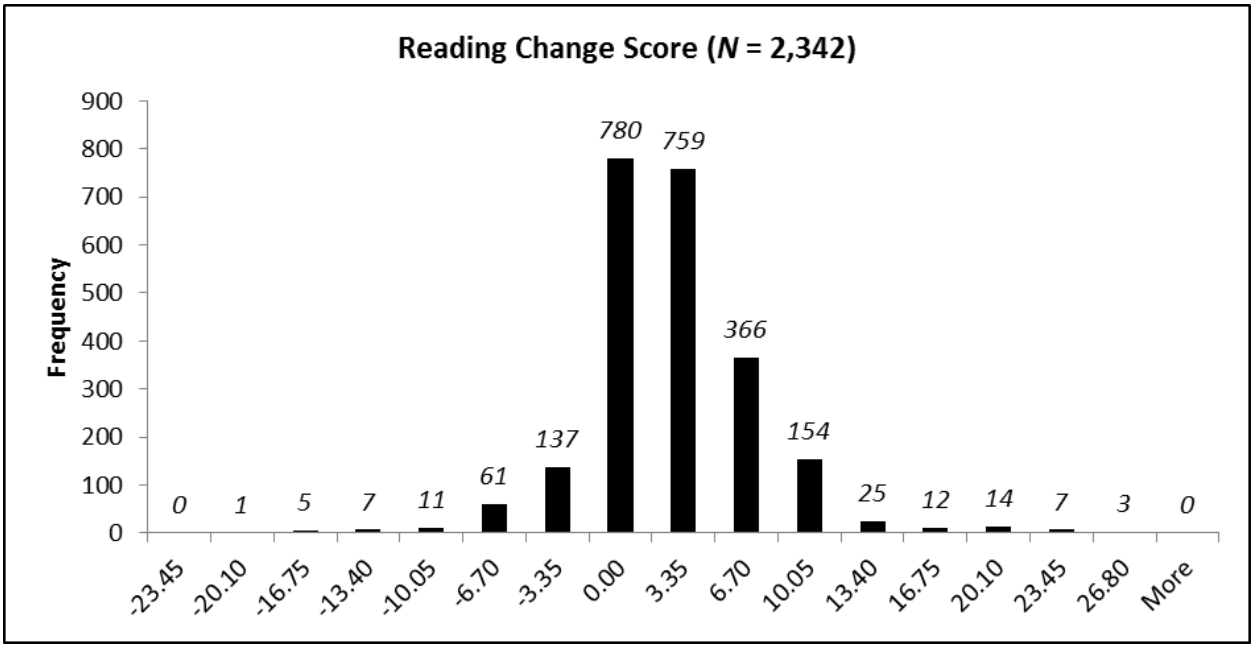

Figure 1. Distribution of change scores for reading.

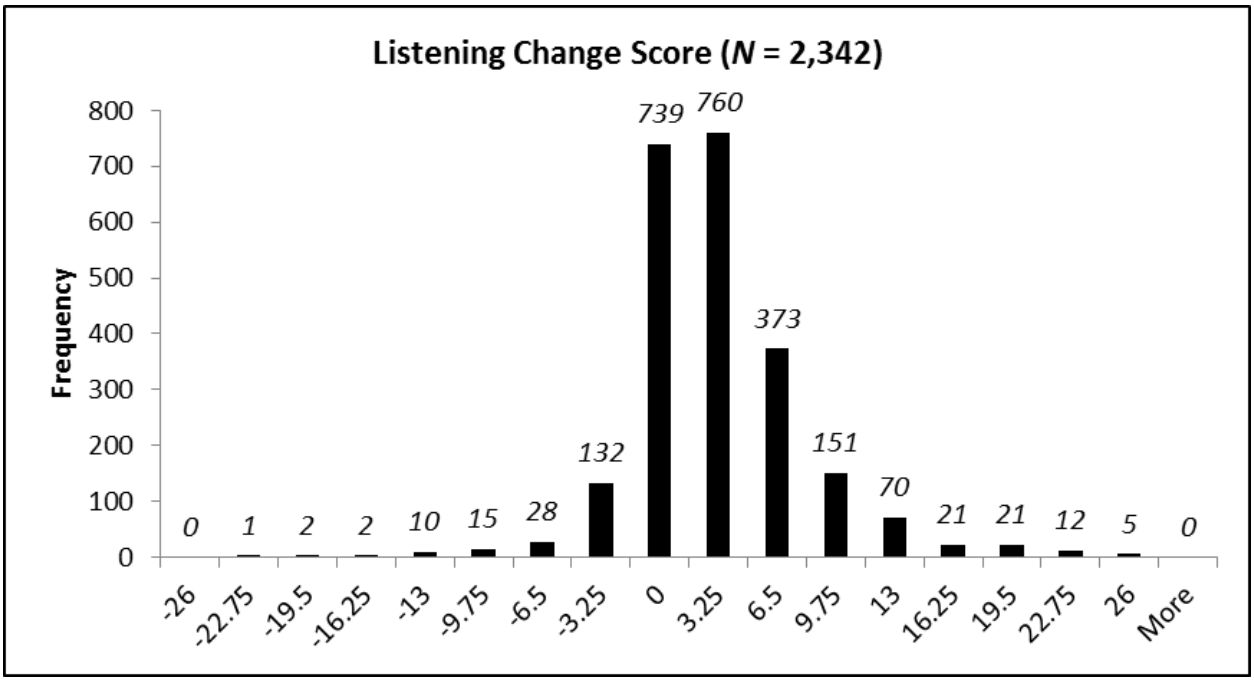

Figure 2. Distribution of change scores for listening.

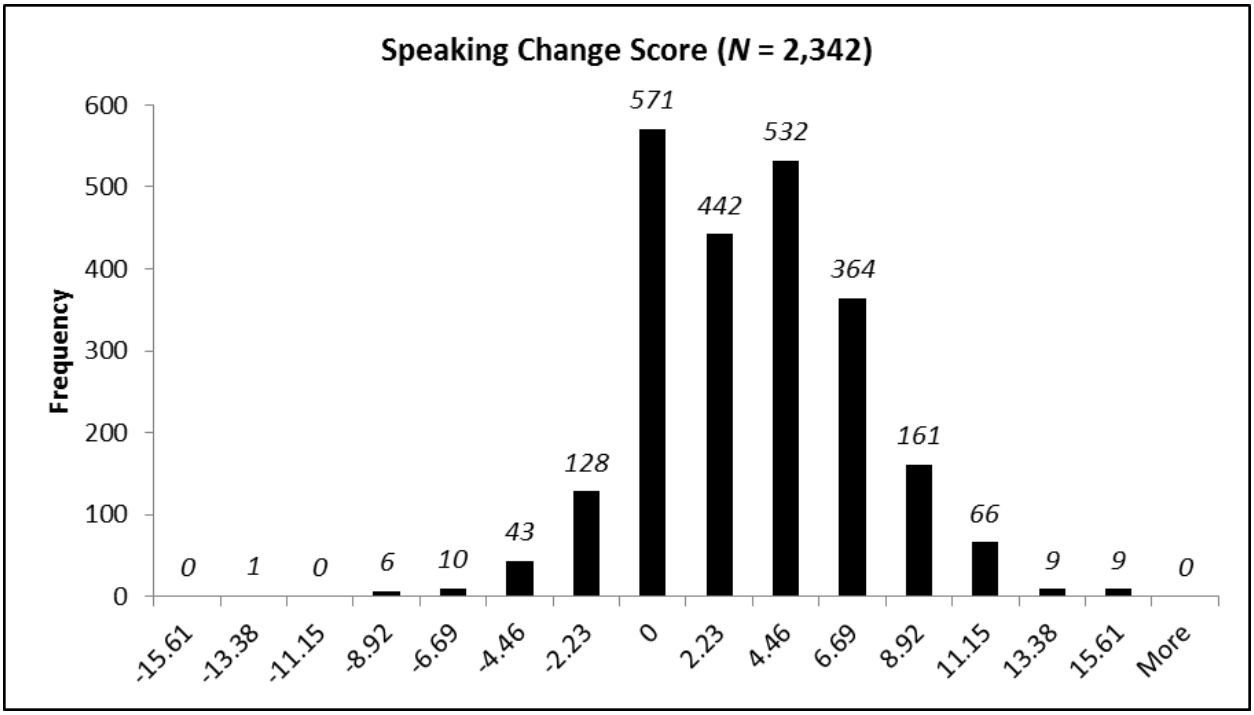

Figure 3. Distribution of change scores for speaking.

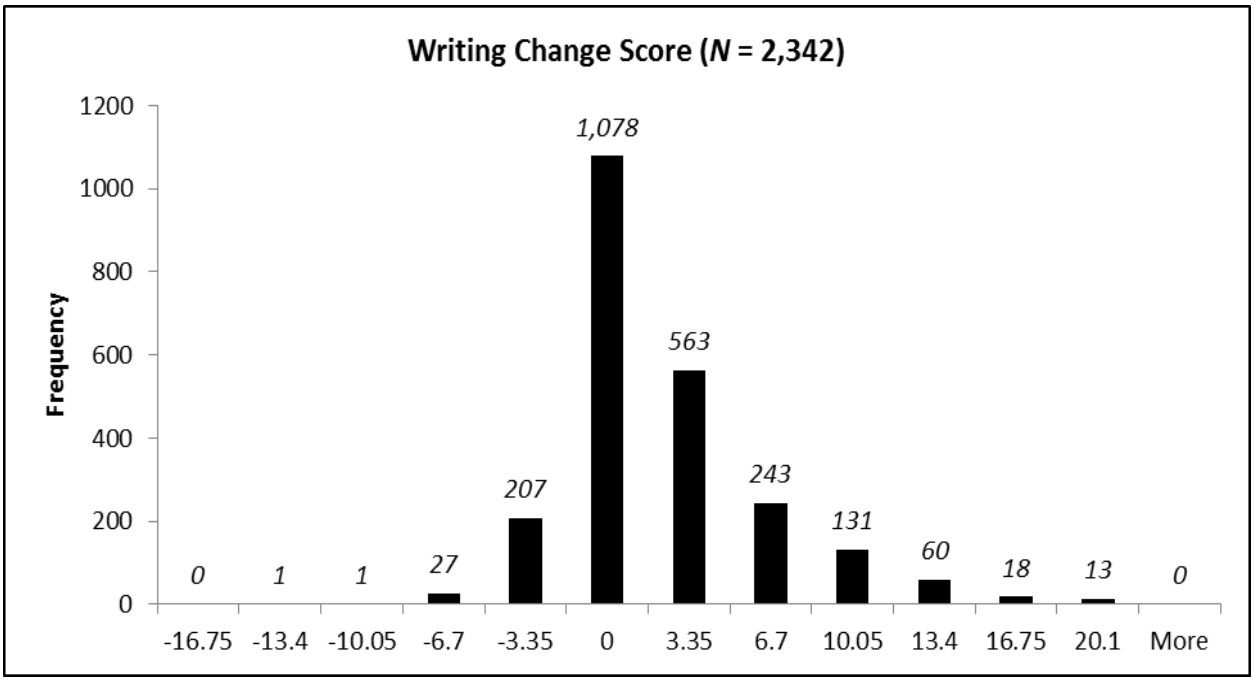

Figure 4. Distribution of change scores for writing.

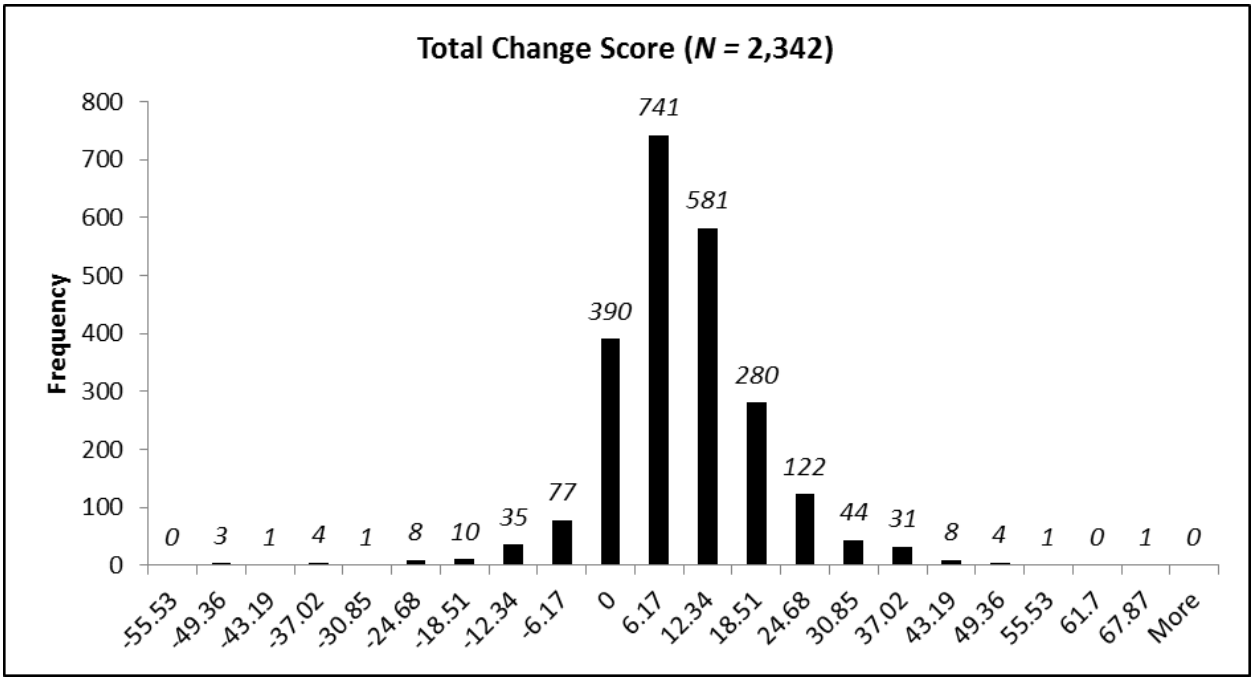

Figure 5. Distribution of change scores for the total test.

We observed at the test level that approximately 77% of the test takers achieved a better performance on the TOEFL iBT test than on the TPO test. The percentages of test takers whose performance improved when taking the TOEFL iBT test were 57% for reading, 60% for listening, 67% for speaking, and 44% for writing. It is worth pointing out that the writing scores for more than half of the test takers actually dropped from the practice test to the high-stakes test, although the average writing change score was positive with a small effect size.

To summarize, on average test takers scored higher on the TOEFL iBT test than on the TPO test. The results also showed that average score differences varied across sections.

Repeated-Measures ANOVA

A series of repeated-measures ANOVA analyses were conducted to examine the effect of the within-subject factor of test consequence. All tests were statistically significant, and the results are reported below.

The mean reading score on the TPO test (M = 22.04, SD = 6.43) was significantly lower than that on the TOEFL iBT test (M = 23.34, SD = 5.65), F_{Reading} (1, 2341) = 190.39, $p < 0.01$, partial eta squared (η^2_{partial}) = 0.08. The η^2_{partial}, a measure of effect size, indicated that the effect of test consequence by itself accounted for 8% of the within-subject variance in the dependent variable.

The mean listening score on the TPO test (M = 22.39, SD = 6.52) was significantly lower than that on the TOEFL iBT test (M = 24.15, SD = 5.05), $F_{\text{Listening}}$ (1, 2341) = 318.56, $p < 0.01$, $\eta^2_{\text{partial}} = 0.12$.

The mean speaking score on the TPO test (M = 21.08, SD = 2.98) was significantly lower than that on the TOEFL iBT test (M = 23.38, SD = 3.75), F_{Speaking} (1, 2341) = 995.29, $p < 0.01$, $\eta^2_{\text{partial}} = 0.30$.

The mean writing score on the TPO test (M = 22.23, SD = 5.48) was significantly lower than that on the TOEFL iBT test (M = 23.16, SD = 4.15), F_{Writing} (1, 2341) = 112.74, $p < 0.01$, $\eta^2_{\text{partial}} = 0.05$.

The mean total test score on the TPO test (M = 87.74, SD = 16.98) was significantly lower than that on the TOEFL iBT test (M = 94.03, SD = 15.97), F_{Total} (1, 2341) = 941.60, $p < 0.01$, $\eta^2_{\text{partial}} = 0.29$.

The results show that the effect of within-subject test consequence was significant for all sections and for the total test. The largest effect size was found for the speaking section (η^2_{partial} = 0.30). At the test level, 30% of the within-subject variance could be accounted for by test consequence. The writing section had the smallest effect size (η^2_{partial} = 0.05). The effect size for listening (η^2_{partial} = 0.12) was similar to that for reading (η^2_{partial} = 0.08).

Generated Linear Model Analyses

In the general linear model (GLM) analyses, the within-subject factor was test consequence. The between-subject factors were interval (short, medium, and long) and mode (timed and untimed). The goal was to determine if there were any interactions between the within-subject and between-subject effects. The results are reported below, and the interactions, if significant, are illustrated in Figures 6 through 9.

For the reading section (Figure 6), a significant interaction was found between consequence and interval, F (2, 2336) = 4.75, $p < 0.01$, $\eta^2_{\text{partial}} < 0.01$. The follow-up analysis of simple effects showed that the performances of all groups improved from TPO to TOEFL iBT ($p < 0.01$). The gain was significantly larger for the medium-interval group ($p < .01$) and the long-interval group ($p < 0.01$) compared to the short-interval group. There was no significant difference in score gain between the medium-interval and the long-interval groups ($p < 0.01$). There was also a significant interaction between consequence and mode, F (1, 2336) = 43.95, $p < 0.01$, $\eta^2_{\text{partial}} = 0.02$. The follow-up analysis of simple effects showed that from TPO to TOEFL iBT, the timed group improved significantly ($p < 0.01$), whereas the untimed group did not change ($p > 0.01$).

Regarding the listening section, a significant interaction was found between consequence and interval, F (2, 2336) = 5.28, $p <0.01$, $\eta^2_{\text{partial}} < 0.01$. The follow-up analysis of simple effects showed that the performances of all groups improved from TPO to TOEFL iBT ($p < 0.01$). The gain was significantly larger for the medium-interval group ($p < 0.01$) and the long-interval group ($p < 0.01$) compared to the short-interval group. There was no significant difference in score gain between the medium-interval and the long-interval groups ($p < 0.01$). There was also a significant interaction between consequence and mode, F (1, 2336) = 18.84, $p < 0.01$, $\eta^2_{\text{partial}} < 0.01$. The follow-up analysis of simple effects showed that from TPO to TOEFL iBT, the timed group improved significantly ($p < 0.01$), whereas the untimed group did not change ($p > 0.01$).

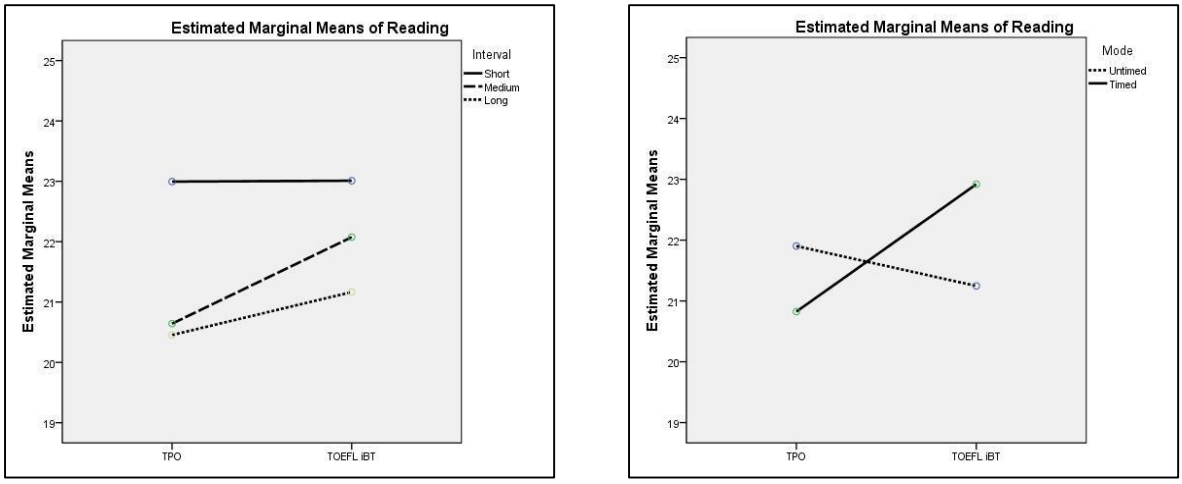

Figure 6. Interaction effects for reading.

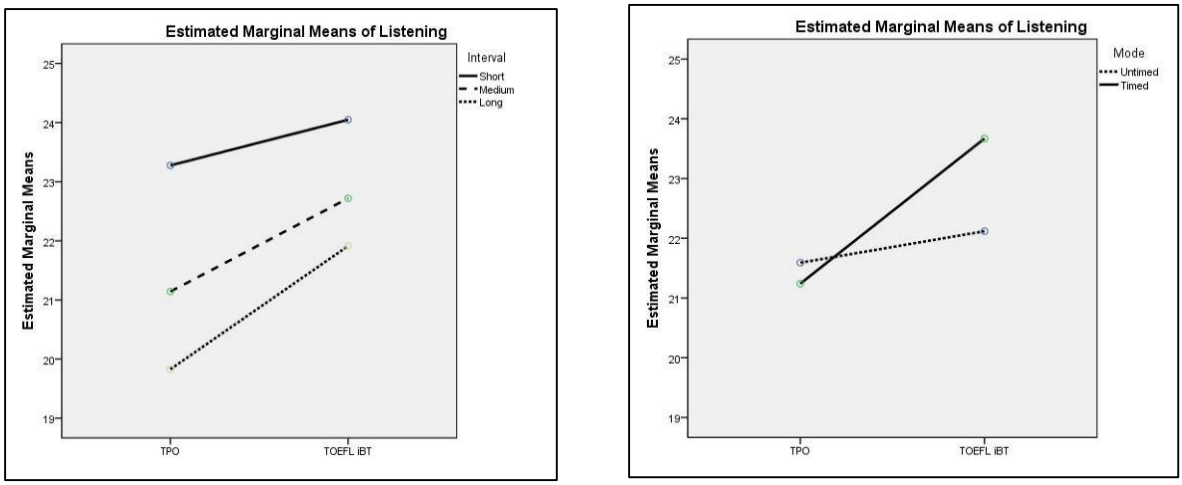

Figure 7. Interaction effects for listening.

With regard to the speaking section, a significant interaction was found between consequence and interval, F (2, 2336) = 8.15, $p < 0.01$, $\eta^2_{\text{partial}} < 0.01$; however, the interaction between consequence and mode was not significant. The follow-up analysis of simple effects showed that the performances of all groups improved from TPO to TOEFL iBT ($p < 0.01$). The gain was significantly larger for the short-interval group ($p < .01$) compared to the long-interval group. There was no significant difference in score gain between any other pairs of groups ($p > 0.01$).

For the writing scores, no interaction effects were found between the within-subject and between-subject factors.

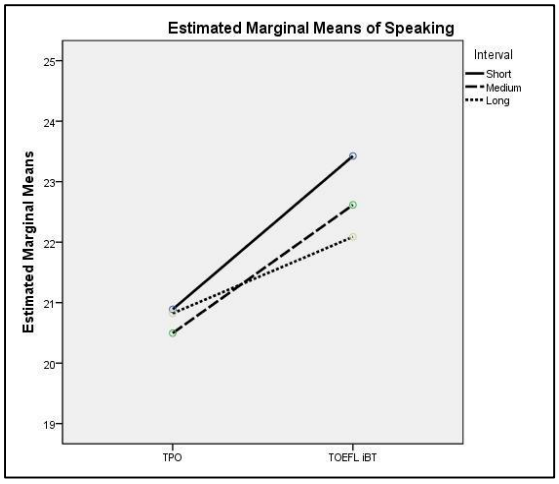

Figure 8. Interaction effect for speaking.

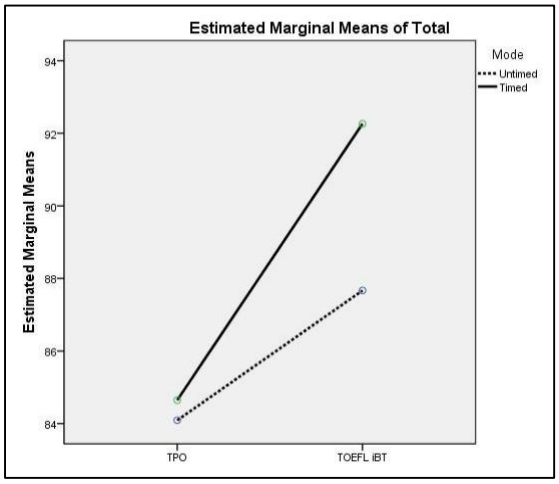

Figure 9. Interaction effect for the total test.

For the entire test, a significant interaction was found between consequence and mode, F (1, 2336) = 19.548, $p < 0.01$, $\eta^2_{\text{partial}} < 0.01$; however, the interaction between consequence and interval was not significant. The follow-up analysis of simple effects showed that the total scores increased from TPO to TOEFL iBT significantly for both the timed group ($p < 0.01$) and the untimed group ($p < 0.01$). The timed group gained significantly more than the untimed group ($p < 0.01$).

Discussion

The goal of this study was to determine if scores on the TPO test can be used to predict subsequent TOEFL iBT test performance. To this end, we examined how test takers performed on the TPO test, a low-stakes practice test, and subsequently, on the TOEFL iBT test, a high-stakes admissions test. In this section, the findings are discussed and interpreted within the framework of expectancy-value theory and in relation to the practical use of TPO scores to gauge test taker readiness for the TOEFL iBT test.

We found that scores on the high-stakes TOEFL iBT test were significantly higher than those on the low-stakes TPO test across all test sections as well as at the test level. This finding is consistent with the expectancy-value model, which predicts that test takers are more motivated to put forward their best efforts and hence achieve better performances on tests of greater consequence than on tests of lesser consequence. However, we do acknowledge the possibility of a practice or instruction effect; that is, the higher TOEFL iBT scores may be in part the result of having practiced on the TPO test or having received additional instructions between the two tests.

We further found that the magnitude of score differences varied across the four skill sections. The tests used in this study consisted of two item types: selected-response reading and listening items, and constructed-response speaking and writing items. The constructed-response items are generally considered to be more cognitively complex tasks, thus requiring more mental efforts from test takers than the selected-response items (DeMars, 2000; Liu et al., 2012; Sundre & Kitsantas, 2004). Previous research (DeMars, 2000; Wolf et al., 1995) found that test takers' affective and emotional reactions to items that are more mentally strenuous are affected by test consequence to a greater degree. Therefore, differential impact of response format on the relationship between test performance and test consequence may be expected.

However, the results from our descriptive and repeated-measures ANOVA analyses were mixed on how response format moderated the relationship between test performance and test consequence. Among the four skills, the percentage of score changes across TOEFL iBT and TPO that were within 2 *SED* was the smallest for speaking and the largest for writing. The score increases from TPO to TOEFL iBT had the largest effect size for speaking and the smallest effect size for writing. Although both sections consisted of constructed-response items, score changes between the high- and low-stakes tests on writing and speaking differed. In comparison to reading and listening, speaking was more affected by test consequence supporting the

expectancy-value theory, but writing was less affected by test consequence, which contradicted the model prediction. We suspect that other factors in the expectancy-value model (e.g., perceived item difficulty in relation to self-estimated ability level) as well as factors that are not included in the model might have contributed to this finding. Differences in scoring approaches for TOEFL iBT versus TPO, which are discussed later in this section, may have also partially accounted for the differential impact of test consequence on speaking vs. writing.

When examining the interaction of test consequence by testing mode, we found significant interaction effects for reading, listening, and the total score. In the cases of reading and listening, the timed group improved significantly from TPO to TOEFL iBT whereas the untimed group did not change. At the test level, a larger mean score difference was found for the timed group than the untimed group. One potential explanation could be that the TPO untimed testing mode provided a less stressful environment and offered more favorable test-taking conditions, such as more processing time, planning time, and response time. These conditions may have helped test takers reduce test anxiety and improve their performance on the TPO test, which consequently reduced the score difference between the low-and high-stakes tests. In other words, the effect of test consequence may have been "neutralized" by testing mode for the untimed TPO test takers. The moderating effect of testing mode, however, was not significant for speaking and writing. One potential reason could be that the extra planning and response time for writing and speaking, and the additional opportunities to process stimulus materials for integrated tasks offered in the TPO untimed mode, may have reduced test takers' anxiety levels but may not have had as much of an impact on the overall quality of test takers' writing and speech because many key indicators of writing and speech quality, such as command of language, pronunciation and intonation, tend to be relatively stable regardless of specific performance conditions. Test anxiety refers to a test taker's emotional reaction to a testing situation and is also part of the expectancy-value model as the affective component. The effect of test anxiety on test performance has been widely examined in educational settings (e.g., Sarason 1980). Our findings showed that the interpretation of test results must consider the interrelatedness of test consequence, the nature and demands of specific tasks, and test anxiety.

With regard to the interaction between test consequence and interval, significant interaction effects were found for reading, listening, and speaking. Test takers in the short-interval group had smaller score gains in comparison to the medium- and long-interval groups on

reading and listening. This could be attributed to the possibility that more learning had occurred during the longer interval periods, giving rise to larger score increases. However, the same pattern did not hold for speaking or writing. The short-interval group had larger score gain on speaking than the long-interval group. We suspect that the short-interval group might have benefited from the practice effect more than the long-interval group. Since the effects of becoming familiar with a test are likely to decrease over time, the longer the interval between retesting, the less the impact on score increase is likely to be. With regard to writing, no significant interaction effect was found. In this study, a long interval was defined as a time period of more than 30 days between the two administrations. The results could imply that it might require longer periods of learning for changes in speaking and writing abilities to occur and for performance gains to manifest in test scores.

Although both mode and interval were found to moderate the relationship between test consequence and test performance for some measures, the size of these interaction effects were small and were essentially not of practical importance. From a practical perspective, it might be reasonable to treat TPO test takers as one population when evaluating the use of TPO results to gauge TOEFL iBT test readiness. Generally, scores across all measures are expected to increase from the TPO test to the TOEFL iBT test. The relatively small effect size of mean score difference for reading (0.21), listening (0.30), and writing (0.19) suggests that it is reasonable to project test takers' TOEFL iBT performance based on their TPO scores for these measures. The relatively large effect size of speaking score difference (0.68) coupled with the weak correlation between the two tests (r = 0.47) warn against using TPO speaking scores to predict test takers' TOEFL iBT speaking performance.

As pointed out in the method section, differences exist in scoring methods for the speaking and writing portions of the two tests. The official TOEFL iBT test utilizes human raters for scoring speaking performance and uses e-rater in conjunction with human raters to score the writing section. In the context of TPO testing, in order to provide instant feedback to test takers on test performance, speaking responses are evaluated by SpeechRater, and their writing responses are evaluated by e-rater without the use of human raters. Although both automated scoring engines are trained to optimally predict human ratings, they do not function exactly like human raters. E-rater more closely emulates the performance of human raters than SpeechRater (Ramineni et al., 2012); however, the human-SpeechRater score correlation was only

approximately .70 for the TPO test with slight increases in more recent upgrades to the engine, indicating a considerable amount of error in the TPO scores produced by SpeechRater (Xi et al., 2008). Therefore, any prediction of TOEFL iBT speaking scores based on TPO test performance should be undertaken with caution.

References

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155−159.

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education, 57,* 119−130.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13,* 55–77.

Eccles, J. (1993). Expectancies, values, and academic behavior. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75−146). San Francisco, CA: Freeman.

Educational Testing Service. (n.d.). *TOEFL iBT research insight series.* Retrieved from https://www.ets.org/toefl/research/ibt_insight_series

Liu, O. L., Bridgeman, B., & Alder, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41,* 352−362.

Napoli, A. R., & Raymond, L. A. (2004). How reliable are our assessment data? A comparison of the reliability of data produced in graded and ungraded conditions. *Research in Higher Education, 45,* 921–929.

Pintrich, P. R. (1988). A process-oriented view of student motivation and cognition. In J. S. Stark & R. Mets (Eds.), *Improving teaching and learning through research* (pp. 55−70). San Francisco, CA: Jossey-Bass.

Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. In C. Ames & M. Maehr (Eds.), *Advances in achievement and motivation* (Vol. 6, pp. 117−160). Greenwich, CT: JAI Press.

Pintrick, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82,* 33−40.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater scoring engine for the TOEFL Independent and Integrated prompts* (Research Report No. RR-12-06). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02288.x

Sarason, I. G. (1980). Introduction to the study of test anxiety. In I. G. Sarason (Ed.), *Test anxiety: Theory, research, and applications* (pp. 3−14). Hillsdale, NJ: Erlbaum.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and nonconsequential test performance? *Contemporary Educational Psychology, 29,* 6–26.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review, 6,* 49−78.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Measurement, 10,* 1−17.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8,* 227−242.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8,* 341–351.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (Research Report No. RR-08-62). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02148.x

Notes

[1] The subgroups were compared to the total sample (N = 2,342) based on the background variables of gender, age, and native language. The results suggested that the subgroups were fairly comparable to the total sample with regard to these background variables.

[2] Standard error of difference *(SED)* is calculated by multiplying the standard error of measurement (*SEM)* by the square root of 2. In the sample, TPO and TOEFL iBT test-taking dates fell between February 2011 and February 2012. The *SEM* estimates listed in the table are the medians across all forms of TOEFL iBT for this time period.

[3.] A mean change score is calculated by subtracting the TPO score from the corresponding TOEFL iBT score. A positive mean change score indicates that the TOEFL iBT score is higher than the corresponding TPO score.