論　　文

Exploring Discrepancies in the TOEFL iBT Scores of Repeat Test Takers

Denise NORTON
Doshisha Women's College of Liberal Arts
Faculty of Liberal Arts
Department of International Studies

Abstract

Students choosing to study abroad either independently or as a component of a local degree must first demonstrate their language proficiency by achieving the required score of their destination institution in a TOEFL iBT or IELTS examination. With such high stakes, many candidates opt to repeat the test a number of times before the submission date in the hope that one of the tests will yield a sufficiently high score. This study analyzed the test results of twenty-five students who, toward the end of a year of preparation, chose to take the iBT TOEFL test two or three times within a month. The aim was to discover whether, and under what conditions, such repeated test-taking resulted in higher scores. Students were also asked in interviews what factors they believed contributed to higher or lower scores. The findings indicated that repeat test-taking was a successful strategy for improving scores, particularly in the reading and listening sections of the test. Students reported factors of format (particularly inclusion of experimental questions), content, environment and test familiarity as contributing to higher/lower scores in repeat tests.

1. Introduction

The internet-based Test of English as a Foreign Language (TOEFL iBT) was introduced in the United States in 2005 and has now replaced the TOEFL computer-based test (TOEFL CBT) as the required test for the admission of non-native English speakers into university programs in the United States. It is also seeking − and to some extent gaining − greater acceptance in countries that have, until now, primarily used the IELTS test as the device for evaluating the suitability of candidates to enter an English-language tertiary program. Like all university-entrance examinations, the new TOEFL iBT is a high-stakes test; consequently, it is not uncommon for students to take the test a number of times in order to achieve a desired entry score. In an ETS study of approximately 250,000 TOEFL iBT candidates over a period of eight months in 2007, Zhang (2008) found that around 10% of candidates repeated the test in that time, with nearly half of these (4.8% of candidates) repeating the test within a month. Zhang's analysis of repeater performance found a moderate to high correspondence between the two tests, with an overall mean score change of +3.74; however the standard deviation of the mean score change was 9.5, indicating a great deal of variance in the individual score changes (Zhang, 2008).

The impetus for this investigation was the predilection of students enrolled in an undergraduate degree requiring one year of overseas study to take a number of TOEFL iBT tests over a short period of time, with some students taking as many as six tests in two months. Although the cost of so many tests was burdensome, many students felt the return would justify the expense, particularly as higher scores were a factor in the award of scholarships that would pay for the cost of the

student's year of overseas tuition. This investigation aimed, first, to explore factors affecting high test score variation in order to determine whether frequent test-taking paid off in higher scores for our student cohort and, secondly, to determine if there were any factors that would indicate when or if repeated test-taking might be indicated.

2. Method

2.1. The candidates

In 2010, ninety-three first-year students enrolled in the International Studies department of Doshisha Women's College of Liberal Arts. The four-year degree program required students to study for one year at an English-speaking university abroad, with the overseas component beginning after the first three semesters had been completed. Applications to the overseas partner institutions needed to be finalized by the end of the first year of study. There was a wide range of partner institutions involved, each having its own requirements in terms of TOEFL iBT scores for entry into award classes. In addition, most institutions offered an ESL or a bridging program (with a lower entry requirement) that allowed students to study an ESL program for the first six months before gaining admission to the award course subjects in the second semester of their overseas stay. The median entry score for the ESL programs was 61 with the median entry to the award course being 80 (out of a possible TOEFL iBT score of 120 divided into 30 points each for the speaking, listening, reading and writing sections).

Students were required to sit a TOEFL iBT test in June of 2010 (during their first semester of study), and another at the end of the year (toward the end of their second semester). The results of this second test determined their choice of overseas university placement. It did so in two ways. (1) Overseas partner universities set their own TOEFL iBT entry requirement to their ESL, bridging or award course classes. (2) Students were ranked internally according to their scores, with higher scores earning students first choice of destination institutions. As well, high-value scholarships were awarded to eight students who performed well in both their TOEFL iBT tests and their grade point average. These scholarships paid the full cost of the student's overseas tuition. With such high stakes, students were highly motivated to perform well. There were no restrictions on the number of iBT tests students could choose to take; they were required only to submit an official score report by January 2011. Many students chose to sit for a number of tests as subsequent tests did not invalidate prior ones.

It was decided to limit this investigation to the twenty-five students who took two or more tests within a four-week period in order to reduce as far as possible the effect of a real increase in language skill, and to include for analysis only those tests that fell within four weeks; however, it should be noted that all of the twenty-five students had prior experience of the test. All students had undertaken the test on at least one previous occasion (in June) and almost a third (7 students) had also taken a test in the month preceding this study. As well, many students went on to take more tests in the month following the end of this study. The results of these prior and subsequent tests were not included in this study.

During the period covered in this study, 14 students repeated the test just once and 11 repeated the test twice within a four-week period. Since students could choose the date and place of their test, individual test dates varied, but all tests were undertaken between mid-November and mid-January and the second (or in some cases the third) tests were completed within 4 weeks of the first.

First-year students were divided into eight different skill level classes based on internal tests on admission to the college. This investigation focuses only on those students who were in the middle four levels. This was a result of student self-selection rather than an intention of the investigation, as it was these middle-level students who were most likely to repeat the test. This was probably due to the level of their scores, which fell between or near the two threshold scores of 61 and 80.

2.2. Test-familiarity and test-wiseness

The tests included in this investigation were taken toward the end of a thirty week (two semester) preparation program with a strong TOEFL iBT emphasis, during which students received nine hours a week of skills classes divided along TOEFL iBT lines into 90 minutes per week each of Integrative Speaking, Integrative Writing, Academic Writing, Public Speaking, Intensive Listening and Intensive Reading. Instructors of the components varied in their approach from those that favored a more test-based syllabus to those that focused more on the underlying skills involved; however, all students were exposed to multiple test-preparation and sample test activities for each of the components of the TOEFL iBT test. As well, all students had taken at least one full practice test under test conditions and one official test (in June). Susser (2010a) discusses the difference between test familiarization and test-wiseness, and the ethicality of the latter versus the former. He quotes Messick (1996) as distinguishing test-wiseness as learned 'strategies that might increase test scores without correspondingly improving skills', contrasting this with familiarization that may 'actually improve validity' (Messick, 1996, p 246). In practice, it seems difficult to become familiar without becoming at least a little wise. In any event, the candidates in this study could be considered thoroughly familiar with the test format prior to taking the first of the tests in this study.

2.3. Interviews

Twenty-four of the twenty-five candidates were interviewed. They were asked: (a) what factors they believed affected their test results; (b) whether some tests seemed easier or more difficult than others.

Table 1: *Score changes (Test 2-Test 1) for each candidate*

	Reading	Listening	Speaking	Writing	Total
	−8	−1	2	−2	−9
	1	4	2	−1	6
	2	10	0	2	14
	1	8	−2	0	7
	−1	6	3	−1	7
	6	−5	0	−1	0
	12	0	0	6	18
	8	−9	1	5	5
	−5	0	0	−3	−8
	5	3	1	−3	6
	10	0	0	6	16
	2	−3	1	3	3
	0	−9	−2	1	−10
	−2	−1	0	0	−3
	3	−6	−2	0	−5
	10	3	0	3	16
	9	6	0	1	16
	−2	1	1	2	2
	−2	−3	−1	0	−6
	4	5	−3	3	9
	3	−7	1	2	−1
	3	4	0	0	7
	−1	−1	3	4	5
	1	1	1	4	7
	0	4	0	3	7
Mean	2.36	0.40	0.24	1.36	4.36
SD	4.86	4.97	−1.48	2.56	8.14

3. Results

3.1. Test Results

Table 1 shows the differences in scores for each of the 25 candidates (calculated as Test 2 score minus Test 1 score) as well as the mean score change and standard deviation. Though the study only consisted of twenty-five candidates, the mean score changes reflected those found by Zhang in his study of 12,000 candidates (Zhang, 2008) with the greatest variation occurring in the reading and listening sections of the test, followed by writing and then speaking. It is noteworthy that the speaking and writing sections of the test, which are marked by raters, offered more reliable (that is, less varied) scores than the listening and reading sections.

Of the twenty-five candidates, fifteen achieved a score in the second test that was five or more points higher than the first score. Of these, five scored more than 10 points higher and four scored more than 15 points higher. In contrast, only six candidates achieved a lower score in their second test, with -10 the lowest negative change recorded.

The most significant changes occurred in the reading and listening scores with nine of the candidates scoring more than a five-point difference in each of these sections. In contrast, only two candidates' scores showed a 5-point change in the writing section and none in the speaking.

3.1.1. Score change over three tests

Eleven of the twenty-five candidates took three TOEFL iBT tests within the month. The table below shows their total scores for each test. Of the eleven candidates, ten achieved a higher score in their third test compared with their first, and one student scored the same on the first and third tests. No students scored lower in the third test. The average score improvement was 8 points.

Table 3 shows the mean score change and standard deviation for each section. The mean score change does not clearly show the degree of variation since negative changes can cancel out positive ones — as in the listening component where there was in fact a high degree of variation in the test scores of individual students. The average standard deviation (Table 4) better reflects the amount of variation in candidates test scores for each section, with the speaking section again showing the least variance.

A study for ETS by Stricker and Attali (2010) into the acceptance of the TOEFL iBT in China, Colombia, Egypt, and Germany found that test takers, when asked to agree or disagree with the statement that the iBT gave them an opportunity to demonstrate their ability, responded positively with regard to the listen-

Table 2: *Comparison of candidate's scores over three tests*

Test1	Test2	Test3	SD	T3-T1
74	65	79	7.09	5
61	67	65	3.06	4
45	59	61	8.72	16
52	59	61	4.73	9
47	54	51	3.51	4
60	60	69	5.20	9
46	64	59	9.29	13
54	59	56	2.52	2
60	52	60	4.62	0
56	62	73	8.62	17
40	56	49	8.02	9
		Mean SD	5.94	
		Mean score change		8.00
		SD score change		5.60

Table 3: *Mean score change by test section*

	Mean score change	SD score change
Reading	4.73	4.05
Listening	0.09	3.91
Speaking	0.82	2.75
Writing	2.36	3.11
Total	8.00	5.6

Table 4: *Average standard deviation by test section*

	Mean SD
Reading	3.68
Listening	2.84
Speaking	1.50
Writing	2.10

ing and writing; however, the reading section received only a more moderate acceptance, and the speaking test received a generally negative response with only 62.5% of candidates from China and a low 28% from Germany agreeing with the statement (the other two countries fell between these scores). Yet in this study, scores for the speaking test proved to be the most stable and reproducible. A possible reason for the non-acceptance of the speaking section of the test may have been that the questionnaire was administered before candidates were apprised of their scores.

3.2 Interview Results

Students cited four factors (other than language ability) that they believed impacted on their test scores on any particular test day. These were: the inclusion of experimental questions; the test-room set-up; familiarity with the topic/content of questions; becoming accustomed to the test format.

3.2.1 The dummy question

Most TOEFL tests include 'experimental' or 'dummy' questions (for the purposes of test development and moderation) either in the listening or reading sections. Students do not know which questions are experimental. Twenty-two students cited dummy questions as a factor in test performance with some saying they preferred a reading dummy while others preferred a listening dummy. Students pointed to two ways in which these dummy questions affected their scores. 1) Firstly, the inclusion of experimental questions required students to maintain concentration for a longer period. The TOEFL iBT is a high-intensity test, with students required to race the clock to complete a question. Once the time allotted has expired, a new question with a new topic comes on to the screen. The inclusion of extra questions requires students to maintain concentration longer-forty minutes longer when the experimental questions are in the reading section; twenty minutes longer if they are in the listening. Although students did not mention this as a factor, the inclusion of experimental questions also requires students to engage with more topics. Some students preferred a dummy listening since this was the shorter in terms of overall time. Others found their concentration waned in the face of extra listening questions. 2) Secondly, some students said they were advantaged when repeating the test, as often the same dummy question would appear. Able then to identify it as a dummy, students did not give it the same intensity of focus. Some students felt that it gave them confidence and helped them to relax, as they were able to understand more, having seen/heard the question previously. One student felt that she lost focus and relaxed too much during the dummy question and that knowing the question was a disadvantage for her. The higher the number of tests taken, the more likely it was that students would encounter a familiar dummy question.

3.2.2 Test room conditions

Test room set-up was the second most cited factor in test score variation. Students did not take all the tests in one test centre; the nine students who mentioned this factor all said they preferred a centre that arranged seating so that all candidates were facing a wall, rather than one where students sat in pairs side by side or facing one another. Factors involved in this were 1) the distraction provided by the nearness of the other candidate; 2) noise − particularly when the adjacent candidate was speaking while the candidate was engaged in listening. Noise was also cited as a problem when students were arranged facing walls.

3.2.3 Content

Seven students spoke of familiarity (or lack of familiarity) with the question content as a factor in test scores. Students reported this as happening most often in listening or reading contexts, but it was occasionally also a problem with an essay question that was outside their realm of experience. Two students talked in more detail about the listening test content, explaining that they 'panicked' and 'lost concentration' when they could not understand a keyword.

3.2.4 Familiarity with test format

Students were asked whether some tests seemed

easier than others, or if they felt tests became easier as they did more. Seven students said they became accustomed to the test over time and could therefore concentrate better. Four students said that, though they had felt a particular test was easier or harder, this was not borne out in the results.

4. Implications

The results of this study suggest that repeated test-taking within a short time-period is a useful strategy in gaining high test-scores on the TOEFL iBT test − particularly where candidates have obtained lower than desired (or expected) reading and listening scores. Based on this sample, it appears that repeating the test is less likely to result in higher scores in the speaking section.

One concern that may arise is the ethics of taking tests repeatedly in order to obtain a desired score − or perhaps more importantly from an instructors perspective, the ethics of advising students to do so. This must be weighed against the ethicality of concealing from students information that might assist them in achieving their academic goals.

Rogers and Yang (1996) cite Millman's definition of test-wiseness as 'a subject's capacity to utilize the characteristics and formats of the test/or test-taking situation to receive a high score' (p. 249). Instructing students that they are more likely to achieve a high score if they take multiple tests certainly seems to be teaching the type of test-wiseness that Susser portrays as ethically suspect (Susser, 2010a). Opposing this is the consideration that the increase in scores may be, not a matter of test-wiseness, but of students acquiring the 'familiarity and anxiety reduction' mentioned by Messick (1996, p. 246) as in fact improving test validity (Susser, 2010a, p. 14). Supporting this interpretation are candidates perceptions that score increases were primarily the result of a reduction in interference caused by experimental question inclusion, and by noise and distractions in the test centre. The third factor students mentioned, that of content, is more problematic. Discussing the TOEFL iBT essay prompt, Susser writes that 'It goes without saying that to be fair to all test takers, the questions cannot require knowledge of any specific content, nor should the topic be culturally biased' (2010b p. 49). The listening and reading sections, though not requiring expert knowledge, do have a central topic, familiarity with which gives the candidate an advantage. The effect of lack of knowledge of a keyword cannot be over-estimated. Taking multiple tests can insure against this-but whether this results in a score that is more valid or less is debatable.

The question of validity of language tests, and how validity is understood by the various stakeholders, has been discussed by many writers in the field. This study is not concerned with validity, but with repeatability and therefore reliability. While there is some overlap between reliability and validity, there is also, as Jones (2001) says 'a potential tension between them' (p. 3) since test reliability (or repeatability) is most easily achieved when test questions are similar (which leads to problems of predictability), or where a narrow range of skills is tested (which leads to problems with usefulness). As Cheng (2006) points out, 'Language test scores...are also affected by the characteristics and content of the test tasks' and the 'characteristics of the test-taker. . .' (p. 25). They are also affected by the characteristics of the test environment. All these factors interact with one another -and the more complex the test, the greater the range of factors involved. How a particular student will interact with a particular set of test questions on a particular day in a particular place depends on a complex range of factors including but not limited to her English ability. Perhaps repeat test taking merely eliminates the non-essential, leaving the student with a best possible score − though this still leaves the question of whether the 'best possible' is a reliable or valid measure of ability.

References

Cheng, L. (2006). *Changing language teaching through language testing: a washback study.* (*Studies in language testing: 21*). Cambridge, UK: Cambridge University Press

Jones, N. (2001). Reliability as one aspect of test quality. *Research Notes: UCLES. 4,* 2–5

Messick, S. (1996). Validity and washback in language testing. *Language Testing.* 13(3), 241–256

Rogers, W., & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment.* 12(3), 247－259. doi: 10.1027/1015-5759.12.3.247

Stricker, L., & Attali, Y. (2010). Test takers' attitudes about the TOEFL iBT. *TOEFL iBT Research Report.* ETS. TOEFLiBT-13

Susser, B. (2010a). Researching the test preparation course: What is needed. *Study Abroad N-SIG Newsletter: Ryugaku* 3 (1), 9–20

Susser, B. (2010b). Problems in assessing EFL writing on high-stakes tests: A guide to the research. *Bulletin of Institute for Interdisciplinary Studies of Culture, Doshisha Women's College of Liberal Arts.* 27, 44–62

Zhang, Y. (2008). Repeater analyses for TOEFL iBT. *ETS Research Reports.* ETS. RM-08-05