Does the experimental set affect TOEFL iBT reading performance?

Bongjun Choi

(Seoul National University)

Choi, Bongjun. 2018. Does the experimental set affect TOEFL iBT reading performance? *SNU Working Papers in English Linguistics and Language 16, 1-13.* This study investigates whether including an experimental set and test takers being aware of it influences TOEFL iBT reading performance. Before going into the experiment, the existence and the purpose of the experimental set is established based on several grounds. To discover a certain item or test functioning differently (DIF) in favor of some test takers or test layout that is meant to be comparable, the concepts of the Mantel-Haenszel procedure (Mantel and Haenszel, 1959) is borrowed. Test scores of participants who are designated into the reference group, focal group 1, and focal group 2 were compared but statistically revealed little regarding the effect of the experimental set. The qualitative approach taken after the test, however, exposed some critical issues related to fairness, performance, and ethics. Along with discussions, some implicit and improvable limitations finally conclude the paper. **(Seoul National University)**

Keywords: TOEFL iBT reading, experimental set, differential item functioning, test reliability

1. Introduction

TOEFL is a high-stakes general language proficiency test used mainly for admission for higher education. It consists of four sections, and a possible reliability issue of the reading section will be dealt with in the current paper. TOEFL is known to be a reliable test, that is, a consistent test that will yield equivalent scores over different occasions of the test. The reliability estimate of the TOEFL iBT reading section provided by ETS (2011) is 0.87. Considering the reliability coefficient's value can range from zero (not reliable at all) to one (perfectly reliable), ETS addresses that the TOEFL iBT reading is a reliable test. However, numbers do not tell us everything and many construct-irrelevant,

unexpected error elements may potentially affect the test takers' performance. These elements include inherent variances within the test taker like the psychological state of the test taker, personality, ethnicity, sex, age and so on. Unexpected circumstantial variants like the testing environment being too loud or too cold could also seriously influence test performance. Regarding reading performance, one element that is actively discussed in the literature is familiarity. For instance, Brown (1982) designed a reading comprehension test consisting of three engineering passages and found out that students majoring in engineering outperformed students of different major background. This was the case not only in specific engineering knowledge, but also in general engineering knowledge. Also, the effect of cultural knowledge was experimented in Keshavarz, Atai, and Ahmadi (2007)'s research. 240 Iranian students participated in the experiment and some were given a reading comprehension test in which the passage was about an Islamic religious leader and others were given passages describing the life of a non-Islamic religious leader. Results showed that the content familiarity was closely related to test scores. In a similar study, Liu (2011) investigated whether a test takers' major field of study and cultural familiarity had an impact on his or her reading performance. Her argument was based on the change from the computer-based test (CBT) to the internet-based test (iBT) that reduced the number of passages and increased the passage length. Since fewer passages (three) appear, topic variety would decrease and possibly favor certain test takers who are familiar with the topic in terms of major field of study or cultural background. Her experiment, however, revealed that most of the items in her test served fairly to test takers with different major and cultural background, even though the content was major or culture specific. To detect this, she used differential item functioning (DIF). According to Zumbo (2007), differential item functioning (DIF) is "a statistical characteristic of an item that shows the extent to which the item might be measuring different abilities for members of separate subgroups." To

say it differently, an item exerts DIF if it is biased. Some concepts of a specific method of measuring DIF will be used in this paper as well. The test takers' native language has also proved to have an impact on TOEFL scores. Alderman and Holland (1981) proved by experiment that certain language proficiency-irrelevant similarities and differences between the test takers' mother tongue and English can act as an advantageous or disadvantageous variable in an English proficiency test performance. Similarly, in Schmitt (1988)'s study, test scores of Hispanics and whites in the U.S. were compared. Both groups reported that English was the language they could perform best in. She discovered that items including Spanish-rooted words functioned favorably to the Hispanic test takers. Other than influences that come from inside the test taker like familiarity, native tongue, and ethnicity, the testing environment could also affect performance. For example, Ling (2016) saw if a specific keyboard type affected TOEFL iBT writing performance. She compared test scores in different environments world-wide; test centers where the U.S. standard English keyboard (USKB) was used and test centers where a country-specific keyboard (CSKB) was used. She found out that keyboard type has little or no impact on test takers' TOEFL iBT writing scores. Since TOEFL iBT is a computer-mediated test, how texts are physically displayed on screen is another important matter. Mary (2004) compiles research related to just this, highlighting that line length, number of columns, window size, and interlineal spacing affect reading time and fixation.

As such, numerous unexpected features can affect test performance and many possible test taker-internal, test taking circumstantial influences are actively being questioned in the literature. However, little attention has been paid to the possible impact of the unscored experimental set. This is perhaps because explicitly establishing the existence of this set in the first place is not convincing owing to ETS not officially showing their stance regarding this matter. In this paper, the potential influence that the unscored, experimental reading set has on the TOEFL iBT reading

performance is investigated based on the argument that this experimental set in fact does exist. Hence, the next section discusses the existence and purpose of the experimental set beforehand as a building block of this paper.

2. Background

The TOEFL iBT is a four-hour test that starts with the reading comprehension section. The reading section may be sixty minutes or eighty minutes depending on how many sets (passage plus questions) are given. The sixty-minute reading test is made up of three sets, while the eighty-minute test consists of four sets. There is nothing special officially said by the ETS about this fourth additional set. The community of test takers and teachers, however, take it for granted that this fourth set is an experimental set of which the performance will not be reflected in the test score. This argument is based on several grounds, the first being the fact that the exact same passage and questions reoccur over different occasions of tests. The crucial point is that the content and questions of the other three sets change, while that of one specific set remain identical. Second, performance on this specific set seemed not to influence test scores. While there is no experimental, published evidence for this reasoning, it relies on qualitative data of individuals, unofficial comments of some test takers and teachers who teach for TOEFL, shared in the community. Some teachers claim that they took the test and did not answer to the questions of the familiar, identical set of the reading section but got equivalent scores to tests including three sets. Finally, constant, substantial amount of attention is paid to this experimental set. In one major Korean TOEFL test taker community, a simple search for “dummy”, which is what the experimental set is also known as in the community (because it is known not to be scored) yielded 16,301 posts from the year of 2002 to 2018.This number is excluding the occurrences

of the word appearing in the comment section.[1] Test takers who are sensitive to test scores would not pay so much constant attention to something that is non-existent in the test. The shared knowledge in the community of test takers is something that cannot be ignored, and it implies that many people who prepare for TOEFL in Korea are aware of the fact that one of the four sets in the TOEFL reading section would not be scored.

Having the claim and three grounds stated, the current study moves on to discuss the purpose of the experimental set. As the word suggests, this set experiments the possible error elements of the passage and questions. Said differently, it is a beta testing process. It checks the smallest things as typing errors to major issues as item difficulty, discrimination. After a number of experimenting, this set is eventually used as a scored set.

The central objective of this paper, then, is to find out if this experimental set affects the reading performance of test takers. If it does, it raises a critical question to the reliability of the TOEFL iBT reading section since this is a systematic issue inherent in the test, not a random, uncontrollable error element. Essentially, test time would be systematically different among test takers, leading to fairness issues. Considering that TOEFL is a high stakes four-hour test, including a twenty-minute unscored section could be psychologically demanding for certain test takers. Moreover, given that test takers are of all ages with different attention spans, this additional twenty minutes may significantly disfavor particular test takers. Another issue dealt with in this paper is the test takers' awareness of the experimental set. Knowing that one-fourth of one's careful effort will not be reflected in their score could be discouraging to some test takers.

Although not handled in this paper, awareness of the experimental set can also be influential in the listening section. If the test takers know that a certain listening set is experimental, they could mute the headphone

[1] The search was done on December 14, 2018. The name of this community is "Gohackers" and the exact search term was "dummy" in Korean.

volume and listen carefully to other test takers speak for the independent speaking tasks and figure out the topic in advance. This is clearly cheating, but there is nothing that the supervisors can do even if they find out because it is up to the test takers not to solve questions. This is also an inherent, systematic issue that could possibly damage the reliability of TOEFL iBT listening section.

Hence, it is worth experimenting whether including the experimental set and test takers being aware of it could, in some way, influence their performance. These two features are studied by an analysis comparing test scores of test takers in different groups according to different test layouts including and excluding an experimental set and different instructions notifying and not notifying the existence of an unscored set.

3. Methodology

3.1. Procedure

To investigate the performance difference of three groups, the concepts of the Mantel-Haenszel procedure (Mantel and Haenszel, 1959) is adopted. Participants are designated into the reference group or the focal group 1 or the focal group 2. The performance of the reference group is used as a guideline to compare the performance of the focal groups. The focal groups 1 and 2 are the groups of interest. Participants designated to the reference group are given tests without the experimental set, while participants of the focal groups are given a test including the experimental set. Only the participants of the focal group 2 are instructed that one of the four sets will not be scored. Score comparison among the three groups is done by analysis of variance (ANOVA). First, comparing the reference group and the focal group 1 gives implication to whether including the experimental set and thus differentiating the test time has an effect on performance. Second, the comparison between the reference group and the focal group 2 hints whether including the experimental set

and raising the test takers' awareness of it influences test performance. Lastly, comparing the focal group 1 and the focal group 2 informs whether raising the awareness of the experimental set has an impact on performance. In this last case, the focal group 1 plays the role of the reference group. The focal group 1 is thus a focal group when compared to the reference group, and a reference group when compared to the focal group 2. An important assumption of the Mantel-Haeszel procedure is the comparability of the participants. According to Ryan and Bachman (1992), "detecting DIF with the Mantel-Haenszel procedure is based on the notion of comparing item functioning for comparable group members". Here, comparability means that the group members should be matched on the qualities that are related to the performance on the item. Participants in this study are matched on their TOEFL iBT reading section test scores.

After the score comparison, a qualitative approach is taken in the form of a short-answer questionnaire. Participants are told which question they got right and wrong along with their total score. The participants are then told about the existence and the purpose of the experimental set. They are then asked to freely fill in a short-answer questionnaire asking their opinion regarding fairness, motivation, and performance.

3.2. Materials

The TOEFL iBT reading test was replicated as equivalently as possible. Passages and questions were those provided by ETS as sample reading tests. The test was reduced to gather as many participants as possible since the original sixty to eighty-minute test was considered too long. Reduction was necessary because the current study is preliminary, giving insight to further study. No compensation was given to the participants for closely solving a twenty-minute reading comprehension test, so their motivation would be significantly lower compared to when they take the actual TOEFL test. To minimize the performance difference stemming

from this motivation gap, the sample test was reduced to one-third. The reduction was done in terms of (i) test time, (ii) passage length, and (iii) number of questions. How the test was reduced is summarized in Table 1. The test time of the real TOEFL reading section is sixty minutes if there is no experimental set. This means twenty minutes is allocated per set, so participants were given a total of twenty-one minutes for the total test without the experimental set. The average reading passage length of TOEFL was approximately 700 words, so 235 words would be an ideal reduction, but reducing the passage to 235 words inevitably damaged the overall coherence and the logical development of the passage. The resulting average passage length is thus 370 words. In terms of question numbers, the actual TOEFL reading set contains thirty-six to forty-two questions (without experimental set), with ten question types, namely, 1. Inference, 2. Vocabulary, 3. Reference, 4. Purpose, 5. Factual information, 6. Negative factual information, 7. Essential information, 8. Sentence insertion, 9. Completing a summary, 10. Completing a table. Completing a summary and table requires close reading of the whole passage, so it was excluded from the reduced version of the set. All the other types of questions were included in the fifteen-question test. In the test that included the experimental set, an additional seven minutes was given, with another 370-word passage containing five questions.

Table 1. Reduction of the TOEFL reading section

	TOEFL reading test		Reduced test	
	Without experimental set	With experimental set	Without experimental set	With experimental set
Test time (per set)	60min. (20min.)	80min. (20min.)	21min. (7min.)	28min. (7min.)
Average Passage length	700 words		370 words	

Number of questions (per set)	36-42 (12-14)	60-66 (12-14)	15(5)	20(5)

3.3. Participants

Nine undergraduate and graduate students participated in the study. The comparability of them was assured by matching them according to their TOEFL iBT reading score. They all received scores ranging from twenty-two to twenty-eight, which is marked 'high' in the TOEFL score report. There were three males and six females. Their age ranged from twenty-three to twenty-six (m=24.9, SD=1.1). They were grouped randomly into the reference and focal groups.

4. Results

Table 2. Raw score, mean, standard variation of the three groups

	Reference group	Focal group 1	Focal group 2
Raw score	P1: 10, P2: 11, P3: 15	P4: 11, P5: 12, P6: 9	P7: 11, P8: 12, P9: 10
Mean	12	10.7	11
Standard deviation	2.6	1.5	1

The reference group seems to have performed better than the focal groups but statistically comparing the score of the reference group and each focal group yielded no meaningful difference (P=0.695, P=0.810). Also, the comparison between the two focal groups also showed no meaningful difference (P=0.976). A further study with more participants per group is needed to quantitively find out if performance is influenced by the experimental set. The results of the qualitative approach in the

form of an interview is discussed next.

As stated above, participants were given feedback regarding their performance on individual items and total score after the test. The participants were also told about the existence and purpose of the unscored set and which set was unscored. They were then asked to voluntarily fill in a short-answer questionnaire questioning their stance regarding fairness, motivation, and performance related issues of the test. Participant 5's response was notable because she got a perfect score for the experimental set, but three questions wrong for the scored sets. She raised some fairness issues because with the same amount of effort, other participants who might have got all the questions for the experimental set wrong could get the same score as she did. Participant 4 answered that he was not able to concentrate on the first two sets, but incrementally became focused and solved questions for the last two sets more efficiently. He actually performed better for the last two sets, the last of which was the experimental set, and said that it is a bit disappointing because he thought he could get a higher score based on his performance of the last two sets. Participant 8 did not know until participating in this study that there is an experimental set in TOEFL iBT reading section and suggested an ethical issue. He asserted that if some questions are unscored, ETS should inform test takers about it and obtain their consent before giving the experimental set to random test takers. He used a drastic expression that it was as if he was being exploited as a "lab rat" for being part of an experiment without having agreed to participate in it.

5. Discussions

Adopting the Mantel-Haenszel procedure to find out whether certain items or a test function differently according to different test layout and instructions revealed no statistically meaningful results. In specific, comparing the scores of the reference group (given test without the

experimental set) to the scores of the focal groups (given test with the experimental set) did not show a meaningful statistical distinction. However, the participants' response in the short-answer questionnaire raised some issues related to fairness, performance, and ethics. The amount of effort a test taker puts into the experimental set turns into nothing and exhausts them. How the experimental set is arranged in the test would thus be an important matter. For slow starters, an experimental set being placed behind would disfavor them because they would become increasingly efficient in solving questions and perform better in the latter part of the test but one of the latter sets would not be scored. This was the case of participant 4 in this study. Alternatively, for test takers with short attention spans, the experimental set being placed behind would favor them because only the questions he or she answered with high concentration would be scored and some possibly sloppy answers or guesses resulting from the lack of attention in the latter part of the test would not be scored. The case is vice versa when the experimental set is in the first half of the test. Slow starters would be advantageous while test takers with short attention span would be disadvantageous. The ethical issue suggested by participant 8 is also an important issue to think about. Test takers have never agreed to be experimented. Test takers may argue that ETS is evading of ethical responsibility by not officially stating the existence of the experimental set and experimenting test takers without consent.

A limitation of this study that is worth discussing is that the whole study starts on the assumption that an experimental set exists in the TOEFL iBT reading section, as argued based on several grounds. However, not everyone agrees with the assumption because there is no official evidence regarding the existence of the experimental set. The implicit limitation of this study is thus that the argument heavily relies on the shared knowledge among the stakeholder community. Another limitation was caused in the process of replicating and reducing the original TOEFL iBT reading test. The actual TOEFL test was not perfectly replicated, so

the results in this paper cannot properly extent to have an implication of the TOEFL test. The equivalence is doubtful mainly because of the reduction process. As stated above, reduction was judged to be necessary due to the fact that this is a preliminary study. However, the reduction resulted in not being able to test the table completion type items, so whatever the construct and ability this item type attempts to measure cannot be measured in the reduced test. Also, the average length of the passage and item numbers were not reduced in equal proportion while the time was. In other words, the time was strictly reduced while the average passage length and item numbers were generously reduced. Relatively, more items and longer passages with shorter time may have forced this test to be a speeded test. The most critical limitation of this paper to consider next is the number of participants and the matching of the participants. Clearly, the mean score of three test takers do not imply much. Furthermore, the matching of the participants was assured by their TOEFL iBT reading test score, but the range of "high" ability in the score report is decidedly too wide. Without doubt, participants who received twenty-two points and participants who received twenty-eight points are not matched on the characteristics related to solving reading comprehension items. Therefore, possible improvements for further study are (i) properly matching the participants according to a strict norm, (ii) recruiting more participants as to be able to calculate meaningful statistical data, (iii) producing a test that is equivalent to the actual TOEFL iBT reading test.

References

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language.* Princeton: Educational Testing Service.

Brown, J. D. (1982). Testing EFL reading comprehension in engineering English. (Doctoral dissertation, University of California, Los

Angeles). *Dissertation Abstracts International*, *43*, 1129A–1130A.
Keshavarz, M. H., & Atai, M. R., & Ahmadi, H. (2007). Content schemata, linguistic simplification, and EFL readers' comprehension and recall. *Reading in a Foreign Language*, 19(2), 19-33
Ling, G. (2017). Are TOEFL iBT® writing scores related to keyboard type? A survey of keyboard-related practices at testing centers. *Assessing Writing*, 31, 1-12
Liu, O. L. (2011). Do 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
Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *Journal of the National Cancer Institute*, 22(4), 719-748.
Mary, C. D. (2004). How physical text layout affects reading from screen. *Behaviour & Information Technology*, 23(6), 377-393
Ryan, K. E., & Bachman, L. F. (1992). Differential Item Functioning on Two Test of EFL Proficiency. *Language Testing*, 9(1), 12-29
Zumbo, B.D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4,* 223–233.

Bongjun Choi
bazubazu@snu.ac.kr