TESL-EJ Teaching English as a Second or Foreign Language
The Electronic Journal for English as a Second Language
May 2019– Volume 23, Number 1

Topic Familiarity Matters: A Critical Analysis of TOEFL iBT Reading Section

Deniz Toker
Western Michigan University
<denizilker.toker@wmich.edu>

Abstract

The central purpose of this paper is to examine validity problems arising from the multiple-choice items and technical passages in the TOEFL iBT reading section, primarily concentrating on construct-irrelevant variance (Messick, 1989). My personal TOEFL iBT experience, along with my knowledge of assessment issues, urged me to critically analyze this particular section so that stakeholders and test designers might revisit the construct of reading, particularly on high-stakes tests, ensuring that test-takers are neither advantaged nor disadvantaged at the time of the test. I attempted to bring to light, or at least point out, the inevitable consequences of using multiple-choice items and the presence of technical terminology in reading passages in such a critical test by drawing on pertinent research. While the findings of some of the research concur with the primary assertion of this paper, others contrast with it. These different perspectives are discussed thoroughly in order to include two sides of the story. I conclude that investigating the significant ramifications of aforementioned factors is a crucial step in establishing the validity of reading sections of all similar tests serving as "both door-openers and gatekeepers" (Bachman & Purpura, 2008).

Keywords: topic familiarity, topic effect, test-wiseness, construct validity, TOEFL, reading assessment

Introduction

If an international student dreams of studying abroad, particularly in North America or Canada, they have to prove their English language proficiency through a reliable assessment tool. In order for one to realize their dreams, this assessment tool is of paramount importance, as it will either open or close the doors for this prospective undergraduate or graduate student. Considering the high-stakes nature of this assessment, one must find an unfailing institution providing a fair test to measure the language abilities/skills of a non-native speaker of English. At the outset, this might sound like a daunting task; however, thanks to its strong face validity, the TOEFL iBT is the first resort of which apprehensive test-takers can think. It is also

worthwhile to add that "TOEFL scores are accepted by more than 6000 colleges, universities, licensing agencies and immigration authorities in 136 countries" (Alderson, 2009, p. 621). Thus, it has recently become one of the most widely used and accepted proficiency tests around the globe. According to Alderson's (2009) comprehensive review of the TOEFL iBT, the emergence of the test dates back to 1961. The first paper-based TOEFL came into existence in 1964, and the Test of Spoken English (TSE) was first used in the late 1970s. A computer-based version of TOEFL (CBT) started to be administered in 1998, and the final version, TOEFL iBT, came about in the USA, Canada, France, Germany and Italy toward the end of 2005. Since then, it has been in use and come under close scrutiny by the creators of the test, Educational Testing Service (ETS), and independent researchers committed to high standards in assessment. As Bachman and Purpura (2008) pointed out, "in high-stakes situations, test developers and test users must insure that decisions based on test scores are as accurate as possible" (p.461); therefore, the onus is on both test-takers and stakeholders to familiarize themselves with the design and the 'impact' of this test. To this end, I will particularly investigate the validity of the reading section of this test by focusing on the construct irrelevance variances and discussing fairness issues springing from the topics of the reading passages. Although there has been a host of research on the questions that I will raise, the findings are, by and large, controversial and inconclusive (e.g., Clapham, 1996, 2000; Cohen & Upton, 2006; Liu, Schedl, Malloy & Kong, 2009). Besides, I will be dissecting this topic both as a novice researcher and a previous test-taker of the TOEFL iBT through an insider perspective.

Reading is a skill which is extremely difficult to directly assess as it is impossible to get into the heads of test-takers to observe the skills and strategies to which they have recourse while reading a text. For that reason, it is rather critical to define the skills/abilities necessary to deal with a reading task. In addition to that, Green (2014) asserts that "in proficiency assessment, designers need to know which knowledge, skills, abilities are essential in the target language use domain" (p. 98). In other words, language assessments should fundamentally include tasks that are as similar as possible to those that candidates will perform in real-life settings. One should then ask if the test-takers of the TOEFL IBT reading will need the same or at least similar skills that they will tap into in an academic setting. When we look at the types of the tasks in the reading section, we come across two main types: 'reading for the basic idea' and 'reading to learn', which can be regarded as academic language abilities. More specifically, Enright and et al. (2000) organized them around four academic reading purposes to guide the design of the TOEFL reading section: "1. Reading to find information (or "search reading"), 2. Reading for basic comprehension, 3. Reading to learn, and 4. Reading to integrate information across multiple texts" (p. 4). The reading section is comprised of three to five long passages, each with approximately 700 words, and test-takers have to complete the whole section in 60 – 100 minutes by answering 12 – 14 questions per passage; all of which are multiple-choice questions. Moreover, the 'reading to learn' items allow for multiple correct answers through which test-takers can get partial credit if they cannot answer all of them correctly. I believe that all multiple-choice questions lend themselves well to test-taking strategies, as they indirectly measure how well one can reach the correct answer without falling prey to one of the distractors.

In her comprehensive literature review of multiple-choice items in reading assessment, Lee (2011) calls attention to "some concerns about the ability of multiple-choice test scores to reflect students' reading skills mastery because the scores alone do not allow interpretation of the reading process" (p.5). Test-takers are not allowed to construct meaning by themselves

utilizing their own meaning-making tools; on the contrary, they avail themselves of some test-taking strategies to eliminate tricky options in multiple-choice questions. At this point, the vital necessity of other skills in addition to reading strategies becomes noticeable, thereby constituting a source of construct-irrelevant variance (Messick, 1989), which simply alludes to the elements that are not intended to be measured in an assessment. Furthermore, according to Rupp, Ferne and Choi (2006),

> *[D]espite newer types of MC [multiple choice] questions that focus more strongly on higher-level reading comprehension, which can be found on the newer TOEFL or SAT versions, for example, we hypothesize that test-takers frequently segment a text into chunks that are aligned with individual questions and focus predominantly on the microstructure representation of a text base rather than the macrostructure of a situation model. As a result, higher-order inferences that may lead to an integrated macrostructure situation model in a non-testing context are often sup- pressed or are limited to grasping the main idea of a text. (p. 469)*

In order to substantiate its strong validity claims, ETS, the creators of the TOEFL iBT test, has been putting too much effort and time into the research of the TOEFL iBT test. To illustrate, Cohen and Upton (2006) undertook an intensive study, which was funded by ETS, to investigate the different strategies used by test-takers when they answered the single-selection multiple-choice items and the new multiple-selection multiple-choice items. Based on the verbal report data obtained from 32 students (mostly Chinese, Japanese, Korean, and other), the following strategies turned out to be prevalent: reading the options first before going back to the passage, rereading the question, paraphrasing the question, and rereading the portion of the passage again carefully. As Lee (2011) pointed out, "the authors demonstrated an array of strategies that were tailored to specific types of multiple-choice questions" (p. 37). Cohen and Upton (2006) also maintained that those test-takers actively benefited from academic reading and test management strategies instead of falling back on "test-wiseness" strategies. Allan (1992) defines test-wiseness as an "ability to use test-taking strategies to select the correct response in multiple-choice tests, without necessarily knowing the content or using the skill that is being tested" (p. 101). However, the limitations of this study make one seriously doubt the validity of their conclusion. In my opinion, test-wiseness strategies are deeply embedded in some cultures, particularly in the Far and Middle East, where test-taking and test-wiseness strategies are essential to get through the examination-driven education systems. Therefore, they might have used these strategies subconsciously or may not have preferred to mention them intentionally for some reason. More importantly, all the participants in this study were mostly high-proficiency students, so they may not have had occasion to use them.

For instance, Lee (2011) examined the strategies used by Chinese-speaking students when given familiar versus unfamiliar topics in a multiple-choice format reading comprehension test, and the "statistical analysis … showed that test-wiseness strategies were used significantly more frequently by the low-performing group" (p. 98). The last limitation stems from the completion of the test without any time limit. In my own TOEFL iBT experience, I distinctly remember how I eliminated the options that were not mentioned in designated paragraph and read the questions and options first, and then located the related text in the last reading passage in order to beat the clock. Unless one masters such strategies, it is a tall order to finish the entire section on time reading the whole passages to learn first and to answer the questions afterwards. Lee (2011) also emphasizes that "multiple-choice testing is a unique format with several answer

options, and that participants sometimes caught and inferred the main points of the passage through the questions or the list of options" (p. 69). Taking all these into account, it can be concluded that the construct irrelevance variance as a result of the multiple-choice items poses a potential threat to the validity of the reading section no matter how well the questions and items are designed.

In this section, I would like to bring up a different but relevant issue that has the potential to overshadow the validity of the TOEFL reading section: the 'topic effect', a term used by Jennings, Fox, Graves, and Shohamy (1999) to find out if "factors such as the test-taker's interest in the topic, prior knowledge of the topic, the perceived relevance of the topic or the test-taker's opinions concerning the topic may have an effect on the test-taker's performance" (p. 427). Before delving into that, I will briefly touch on my own TOEFL iBT test experience again to give the reader the reason that acted as a catalyst for this research. While I was reading the passages with high stress and anxiety due to the ticking time on the screen, I was overwhelmed by their content and length. What is more, most of the reading passages were about architecture, arches and domes as far as I vaguely remember, of which I had/have very limited knowledge. After the test, I was debating in my head whether I could have done better if the topics had been within my field of interest and could not help thinking about how advantageous the ones that studied architecture were. I do not deny the fact that all the information needed to answer the questions was already provided by the text, as claimed by ETS; however, the technical terminology in texts put a lot of strain on me, thereby taking more time and effort to understand the concepts and retain them in my memory till the last summary question. For this reason, the topics chosen for the TOEFL iBT reading section might be favoring some candidates inadvertently due to their prior knowledge. Therefore, my argument here has something to do with the fairness of the reading section arising from the 'topic effect', which brings to mind the 'construct irrelevance variance' (Messick, 1989) one more time.

First of all, it is a widely accepted fact that the reading process is much more complicated than decoding textual symbols both in one's L1 and L2. The literature review done by Lee (2011) put forward that reading should not be regarded as a simple receptive skill due to the other skills and resources involved in the process. For instance, Radojevic (2006) described reading comprehension as the "process [in which] readers construct a mental representation of the author's message, which includes both the information in the text and its interpretation by the reader" (p. 14). In addition to that, Dechant (1991) claimed that readers construct meaning from the text with the help of their background knowledge. Lee (2011) also asserted that "instead of viewing readers as passive decoders, researchers emphasized the role of readers as they actively engaged in the reading process by the knowledge they brought to the text" (p. 10). Therefore, it is safe to state that the prior knowledge of the readers plays a significant role in reading comprehension. So, if we consider the technical academic passages in the TOEFL iBT reading section, test-takers that are familiar with the topic inherently have an advantage over the others who are not in that they can comprehend and recall the textual information better and faster than the others thanks to the schemata they already have. Observing the effect of prior knowledge, Tsui (2002) maintained that "readers at a lower level of [second] language proficiency could perform better than, or at least as well as, readers at a higher level of language proficiency" (p. 29). Before jumping to a hasty generalization, I would like to share a few more studies that looked into the 'topic effect' in reading tests. Clapham (1996) analyzed the reading performances of Business and Social Sciences students in reading section of the International English Language Testing System (IELTS) test zooming in on performance differences in

understanding texts in and out of their disciplines. She observed that students in general did significantly better on the reading module in their own subject area. However, Clapham (2000) later added:

> *[W]hile lower level students could not take advantage of their background knowledge because they were too concerned with bottom-up skills such as decoding the text, and while high proficiency students were able to make maximum use of their linguistic skills so that, like native speakers, they did not have to rely so heavily on their background knowledge, the scores of medium proficiency students were affected by their background knowledge. (p. 515–16)*

Chung and Berry (2000) also found that linguistic proficiency was a better predictor in reading comprehension compared to the background knowledge of the test-takers when examined the performances of their subjects in the IELTS reading test of science/technology module and a science text. (as cited in Lee & Anderson, 2007). Students with high linguistic proficiency used their background knowledge more effectively in their study. It is also noteworthy to mention the challenges faced in these studies in determining the proficiency levels, the decision of the cut-scores, and the extent of topical knowledge of test-takers. Still, it can be reasonably inferred from these studies that test-takers with intermediate proficiency levels capitalize on the topical knowledge that they have.

I would also like to include the remarks of a principal assessment designer at ETS whom I contacted regarding this particular topic. She thoroughly explained to me how arduous the reading selection process is and how diligently the committee members work on that to ensure the fairness of the reading texts. She also pointed to a study carried out by her and other researchers who concentrated on the topic effect or in their terms, 'content knowledge.' Liu, Schedl, Malloy and Kong (2009) set out to ascertain if content knowledge affects TOEFL iBT reading performance. Although their literature review comprises several studies proving the effect of prior knowledge in reading tests, they conducted a more comprehensive study with a substantial number of previous TOEFL test-takers (n= 8,692). In summary, they concluded that despite the density of physical science topics or culture-related content of the reading passages, the great majority of the items displayed no differential item functioning (DIF), which they used as a method to investigate the impact of prior content knowledge on TOEFL iBT reading performance along with differential bundle functioning (DBF). Besides, they asserted that "the analysis of many of the items displaying DIF suggests that the differences in performance may be construct-relevant differences based on real differences in certain aspect of the language ability that TOEFL iBT targets" (p.18). However, the limitations of their study, such as the disregard of the proficiency level of test-takers and the groups' cultural diversity, cast doubt on the validity of the findings, for a number of "studies suggested that the effect of background knowledge varies according to the level of language proficiency" (Krekeler, 2006, p.100). Finally, it is also worth citing Peirce (1992) who pointed out this issue from a different perspective while demystifying the TOEFL reading section in its old version:

> *[W]hether or not a candidate has background knowledge about real estate prices in the United States, the candidate still has to have sufficient command of' the English language to understand the question in order to answer it correctly. ... If the language of the question is easier than the language of the text, Candidate B would have an advantage over Candidate A with respect to background knowledge and time. However, if the language of the question is no simpler than the language of the text, then the only*

advantage that Candidate B would have over Candidate A would be a time advantage. (p.683)

Despite his plausible explanation, I hold the opinion that the time advantage certainly makes a difference in timed tests like TOEFL iBT. I vividly recall how stressed I was because of that during the test and had to read through the terminology of architecture over and over again to get a solid grasp of it in order to answer the detail questions. Alshammari (2013) notes that "learning to read under timed conditions is a skill that needs systematic practice to be fully developed" (p. 3). His study provides valuable insights into the time constraint on second language reading comprehension. He investigated it recruiting of 47 Saudi participants who were learning English as a second language and had similar level of English proficiency. They were put into three time groups: limited (20 minutes), extended (30 minutes), and unlimited (40 minutes) and were assessed through a reading text adapted from a standard TOEFL test. The statistical analysis of the data acknowledged the fact that time constraint is a significant factor in participants' overall reading scores, which corroborated "the findings of some previous studies which found that the more time participants had, the better their scores on a reading comprehension test were (e.g. Lesaux, Pearson, & Siegel, 2006; Meyer, Talbot, & Florencio, 1999)" (Alshammari, 2013, p.29). Another study lending support to this argument was conducted by Nguyen's (2012) who explored the impact of schemata (background knowledge) and time constraint on the reading comprehension of Vietnamese learners of English as a second language (N=32), and the findings revealed that background knowledge and unlimited time constraint led to the best performance on the reading task. Nevertheless, both Alshammari (2013) and Nguyen (2012) reported that their findings contradict some other studies (e.g. Chang, 2010; Cushing-Weigle & Jensen, 1996), and yet they attributed the discrepancies to the subjects' long training procedures in those studies.

On the whole, I still contend that 'topic effect' or 'content knowledge' might be partly responsible for the overall reading scores in the TOEFL iBT test as it affords test-takers with basic proficiency levels extra time and speed to get through technically-loaded passages, which, in return, threatens the validity of the entire reading section. I must also recognize the fact that it is one of the variables over which test designer cannot have so much control considering the wide range of test-takers, and yet in academic reading tests, the context should be more neutral to avoid giving any kind of advantage to a specific group of test-takers due to their field of study. In my opinion, exceedingly terminological reading passages, not only in TOEFL iBT but also in any proficiency tests, indirectly involve some other sub-skills as well as reading strategies, and the comprehension of advanced texts does not happen without basic/prior knowledge of the topic no matter how proficient a nonnative speaker is. Moreover, in real life situations, we do not have to read such long texts in a limited time maybe except for examinations, even then we are supposedly familiar with the subject or informed about it in advance.

In this critical analysis, I have endeavored to analyze one specific section of such a reputable standardized test, TOEFL iBT through the lens of validity and fairness highlighting the potential construct irrelevance variances. Many years ago, Elwein et al. (1988) articulated that "as far as validity is concerned, for the summative purposes of certification, selection, monitoring and accountability that have come to dominate the public face of assessment, face validity is what counts." (as cited in Broadfoot, 2005, p. 133). However, we should not take anything at face value and should regard validity as a multifaceted concept that can be

investigated in several other ways such as content validity, construct validity, or consequential validity. It behooves all institutions and organizations like ETS developing assessment materials to take notice of the concerns emerging from their test constructs, not only to maintain high standards in testing, but also not to allow the legitimacy of decisions made upon exam results to be adversely affected. With regard to that, ETS (2011) acknowledges the fact that "concerns about test validation were an integral part of the test design process" and adds that "test validation is an ongoing process that continues to be actively supported by ETS and the TOEFL Board through the Committee of Examiners (COE) Research Program" (p. 10). So far, I have considerably benefited from the power of scientific research to support my arguments, and yet I would like to end my analysis with a thought-provoking quote by Shohamy (2001) that has deeply affected the way I think of all types of standardized tests and the related research on them:

> *Tests use the language of science. The use of statistics and the presentation of tests and examinations in scientific terms gives the process authority, so that the public make an assumption that testing systems are 'fair' and 'trustworthy' even if they are not. (as cited in Fulcher & Davidson, 2006, p. 150)*

About the Author

Deniz Toker graduated with an M.A. in TESOL from Western Michigan University, where he worked as a graduate research assistant in the Special Education and Literacy Studies department. He earned his B.A. in English Linguistics from Hacettepe University in Turkey and then obtained CELTA. He taught English as a foreign language at various institutions in Turkey for several years. He is currently teaching English as a second language to refugees and immigrants. His main research interests lie in culturally responsive and critical pedagogies and fairer assessment practices for language learners.

References

Alderson, J. C. (2009). Test review: Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®). *Language Testing, 26*(4), 621-631. doi:10.1177/0265532209346371

Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing, 9*(2), 101-119.

Alshammari, H. A. (2013). *Effects of time constraint on second language reading comprehension* (Unpublished master's thesis). Southern Illinois University, Carbondale, USA.

Bachman, L. F., & Purpura, J. E. (2008). Language assessments: gate-keepers or door-openers? In B. Spolsky & E. Hult (Eds.), *The handbook of educational linguistics* (521-532). Hoboken, NJ: Wiley-Blackwell.

Broadfoot, P. M. (2005). Dark alleys and blind bends: Testing the language of learning. *Language Testing, 22*(2), 123-141. doi:10.1191/0265532205lt302oa

Chang, L. Y. H. (2010). Group processes and EFL learner' motivation: A study of group dynamics in EFL classrooms. *TESOL Quarterly, 44*(1), 129-154. http://dx.doi.org/10.5054/tq.2010.213780

Chung, T., & Berry, V. (2000). The influence of subject knowledge and second language proficiency on the reading comprehension of scientific and technical discourse. *Hong Kong Journal of Applied Linguistics, 5*(1), 187-225.

Clapham, C. (1996). The development of IELTS: A study of the effect of background knowledge on reading comprehension. *Studies in Language Testing Series, Volume 4*. Cambridge, UK: Cambridge University Press.

Clapham, C. (2000). Assessment for academic purposes: Where next? *System, 28*(4), 511-521. doi:10.1016/s0346-251x(00)00034-8

Cohen, A. D. & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks*. TOEFL Monograph Series, MS – 33. Princeton, NJ: Educational Testing Service.

Dechant, E. (1991). *Understanding and teaching reading: An interactive model*. Hillsdale, NJ: Lawrence Erlbaum.

Educational Testing Service. (2011). Reliability and comparability of TOEFL iBT scores. *TOEFL iBT Research Insight, 1*(3), 1–8.

Enright, M. K., Grabe, W., Koda, K. Mosenthal, P. B., Mulcahy-Ernt, P., & Schedl, M. A., (2000). *TOEFL 2000 Reading Framework: A Working Paper* (Report No. RM-00-04). Retrieved March 01, 2018, from https://www.ets.org/research/policy_research_reports/publications/report/2000/iciv

Fulcher, G., & Davidson, F. (2010). *Language testing and assessment: an advanced resource book*. Abingdon: Routledge.

Green, A. (2014). *Exploring language assessment and testing: language in action*. Abingdon, Oxon: Routledge.

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers choice: an investigation of the effect of topic on language-test performance. *Language Testing, 16*(4), 426-456. doi:10.1177/026553229901600402

Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: The effect of background knowledge revisited. *Language Testing, 23*(1), 99-130. doi:10.1191/0265532206lt323oa

Lee, H., & Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Testing, 24*(3), 307-330. doi:10.1177/0265532207077200

Lee, J.Y., (2011). *Second Language Reading Topic Familiarity and Test Score: Test-Taking Strategies for Multiple-Choice Comprehension Questions*. PhD (Doctor of Philosophy) thesis, University of Iowa.

Liu, O. L., Schedl, M., Malloy, J., & Kong, N. (2009). Does Content Knowledge Affect TOEFL IBT™ Reading Performance? A Confirmatory Approach to Differ*ential Item Functioning. ETS Research Report Series, 2009*(2), I-29. doi:10.1002/j.2333-8504.2009.tb02186.x

Messick, S. 1989: Validity. In Linn, R.L., editor, *Educational measurement*. New York: American Council on Education/Macmillan.

Nguyen, T. T. T. (2012). *The impact of background knowledge and time constraint on reading comprehension of Vietnamese learners of English as a second language* (Order No. 1529505). Available from ProQuest Dissertations & Theses Global. (1220449702). Retrieved from http://libproxy.library.wmich.edu/login?url=https://search.proquest.com/docview/1220449702?accountid=15099

Peirce, B. N. (1992). Demystifying the TOEFL® Reading Test. *TESOL Quarterly,26*(4), 665-691. doi:10.2307/3586868

Radojevic, N. (2006). *Exploring the use of effective learning strategies to increase students' reading comprehension and test taking skills*(Unpublished doctoral dissertation). Brock University, St. Catherines, Ontario, Canada.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing, 23*(4), 441-474. doi:10.1191/0265532206lt337oa

Tsui, Y. (2002). *Effects of English language ability, vocabulary knowledge, and content familiarity on comprehension performance and strategy use of high school readers learning English as a foreign language* (Unpublished doctoral dissertation). University of Kansas. Lawrence, USA.