An Analysis of the TOEFL

Terra A. Minolli

Monterey Institute of International Studies

An Analysis of the TOEFL iBT

Hearing the word “test” often evokes feelings of anxiety, anger, fear, and may even demotivate learners. The word is hardly ever associated with a feeling of warmth and excitement. While instructing English at two Korean universities, I witnessed firsthand the process of anxiety, frustration and disappointment that arose from the test. I have seen how results can facilitate or diminish motivation and furthermore change an individual’s life. Given this assignment, I saw an opportunity to deeply analyze the TOEFL iBT to better understand the psychological impact that TOEFL iBT test takers experience. Despite inevitable revisions, I hope ETS continues to discover innovative ways in increasing positive washback and limiting negative.

History of TOEFL

In the early 1960’s the Test of English as a Foreign Language (TOEFL) developed to assess language proficiency for non-native English speaking students wishing to study at North American universities. The first test was administered in 1964, and by 1965 the College Board and Educational Testing Service (ETS), a non-profit education organization, took responsibility for developing and administering the test. The original TOEFL paper-based test (TOEFL pBT) evolved to the TOEFL computer-based test (TOEFL cBT) and later to the TOEFL internet-based test (TOEFL iBT). As the format changed, the structure did as well. It evolved from discrete-point and multiple-choice questions to include performance-based and direct testing tasks (Qian, 2010). Discrete-point testing is a test that measures “one and only one linguistic element” (Bailey & Curtis, 2015, p. 65), while a direct test “involves doing the skill involved” (Bailey & Curtis, 2015, p. 64). In other words, the focus of the test shifted from grammatical accuracy to communicative competence.

The TOEFL Paper-based Test

In 1964 the TOEFL pBT was first released and is still used today, however, it is not commonly used as the Internet is now ubiquitous. The first version contained 5 sections focused on discrete-point components (ETS, 2011). In 1976 the developers reduced the 5 sections into 3: listening comprehension; reading comprehension; and structure and written expression (Educational Testing Service, 2010b). It takes about three and half hours, is scored on a scale from 310 to 677, and currently costs 170 dollars (Educational Testing Service, 2014e).

The TOEFL Computer-based Test

As the efficiency of computers evolved, ETS introduced TOEFL cBT in 1998 as a measure to move towards using electronic language testing (ETS, 2007). Although there were minor changes to the listening and reading sections, the test incorporated a new section: writing. Questions on the cBT were computer adaptive – they changed according to how questions were answered. For example, if answered incorrectly, then the following question would be of equal or lesser difficulty. The TOEFL cBT has been discontinued since the development of the TOEFL iBT.

The TOEFL Internet-based Test

The validity of measuring language proficiency concerned administrators and scholars as TOEFL scores of 550 and above did not correlate to students' communicative abilities (Zareva, 2006a). In 2005 the TOEFL iBT launched with drastic revisions. Along with an included speaking section, the content focused specifically on academic English that students would use in every day academic life.

The TOEFL iBT is administered through a secure network at predetermined test centers around the world (Educational Testing Service, 2008)**.** Unlike the TOEFL cBT, the iBT is not

computer adaptive. The test contains 4 sections measuring different language modalities: reading, writing, listening and speaking. The four sections are scored on a scale of 0 - 120. It takes about four hours, including a 10-minute break. Depending on the country the test is administered in, it costs between 160 and 250 U.S. dollars (Educational Testing Service, 2014b). Test registration is by phone, mail or online (see Appendix A for the phone and mail registration form).

The iBT & The 4 Skills

Canale and Swain (as cited in Bailey, 2015) define communicative competence as a sum of four competencies: grammatical, sociolinguistic, strategic and discourse. The TOEFL iBT is a proficiency test for entrance into an academic institution. It is essential that learners can demonstrate professional academic work and communication before entering an undergraduate or graduate program in North America. To test communicative competence encompassing all aspects of academic life, the TOEFL iBT measures proficiency in the four skills. Table 1 breaks down the four skills in detail. Appendix B is a sample test provided free of charge by ETS, also accessible online.

The TOEFL iBT is administered through a secure internet-based network (Educational Testing Service, 2008). All test takers must report to a test center to take the test. Centers are equipped with computers, keyboards, headsets and microphones. Verbal responses are digitally recorded and sent to ETS' online scoring network. Similarly, the responses from the writing section are typed and sent to the network (Educational Testing Service, 2008) To mimic an academic setting, test takers are permitted to take notes through all sections of the test, however, they are collected and destroyed at the end of the test. A persistent "toolbar" tracks time, adjusts volume, and lets test takers go back-and-forth between the reading questions.

The TOEFL Speaking Section

My dream position is to teach conversational English to university students abroad who have intentions to either study or work in North America. Due to this dream, I will focus strictly on the speaking section of the TOEFL iBT for the remainder of this paper.

Unlike its predecessors, the iBT incorporates a speaking section with integrated tasks - tasks exercising more than one skill (Bailey & Curtis, 2015)**.** What drove this inclusion was the realization that understanding coupled with effective communication was the key to success in an academic setting (Zareva, 2006a). Tannenbaum and Wylie (2004) mentioned that in the past university admissions required a 550 score or higher, which was believed to reflect learners' readiness to learn in an academic setting (as cited in Zareva, 2006, p. 46). However, research suggested that this score did not reflect learners' ability to affectively communicate and participate in academic programs (Zareva, 2006a). With validity concerns, the developed TOEFL iBT included a speaking section that assesses realistic academic scenarios. See Table 2 for the speaking task outline.

The Scoring System

The TOEFL iBT is a norm-referenced test in which individual performance is reported in in relation others. Multiple certified human raters score the test using a 1 – 4 rubric (see Appendix C). The tasks are summed and reported on a scale of 0 – 30. The levels are distinguished by *weak* (0 - 9), *limited* (10 -17)*, fair* (18 -25) and *good* (26 – 30). In addition to the scaled score report, test takers receive personalized performance feedback. ("Interpret scores," 2014). The scores are valid for two years.

To ensure scoring reliability ETS: (a) provides detailed specifications to guide raters, (b) provides online practice tests, (c) vigilantly monitors scoring and, (d) provides detailed holistic

rubrics (Tumposky, 2009, p. 5). ETS argues that multiple raters minimize rater bias as rater judgments contribute to the raw test scores (Educational Testing Service, 2014d). In addition, raters follow detailed rubrics for rating the integrated and independent speaking tasks. For security purposes tests are not scored at test centers. They are scored anonymously and objectively through a centralized scoring network.

The holistic scoring method, which assesses students' overall performance on each task, is not problem free (Xi & Mollaun, 2006). According to Douglas and Smith (1997), the most frequently acknowledged problems that ETS arguably recognizes are inconsistencies across raters and calibrating raters for reliability (as cited in Xi & Mollaun, 2006). Colby-Kelly and Turner (2007) argue that holistic scoring systems may be too broad to capture language development (as cited in Jamieson & Poonpon, 2013). Analytic scoring is a possible remedy to holistic scoring. However, this too is not without problems. In Xi and Mollaun's research (2006), they found that raters scored integrated tasks more consistently than independent tasks.

Score reports provide diagnostic information to both test takers and institutions. For test takers, the scaled scores report their level as weak, limited, fair or good. Knowing one's strengths and weaknesses on a test provides guidance for further study. From an admissions perspective, scores are interpreted and used to either accept or decline admissions into an undergraduate or graduate program. In addition, speaking scores may be useful for teachers planning lessons.

ETS provides published guides for all parties involved in the testing process that are readily available to download on ETS' website. The guide for test proctors is included in Appendix D. The "Information for Score Users, Teachers and Learners" report offers general insight for interpreting scores, aimed at all groups (Educational Testing Service, 2011). A copy is included in Appendix E. Though ETS does not set hard requirements for minimum entrance

scores, leaving institution to determine those according to their specific policies, they provide a "Setting the Final Cut Scores" document as preliminary guidance (Educational Testing Service, 2005). A copy is included as Appendix F.

The Reliability and Validity of TOEFL iBT Speaking

By definition a test is "reliable" if it "measures its intended constructs" (Bailey & Curtis, 2015). The TOEFL speaking section measures learners' abilities to communicate in an academic setting. The six tasks from the integrated and independent tasks make the test reliable as it integrates listening and reading skills and has the test taker defend opinions. In a real academic setting students attend lectures, participate in discussions and interact with individuals throughout campus.

Although the test of speaking measures its intended constructs, the overall score reported to institutions is an issue of reliability. Given that "high stakes" decisions are made from section scores, Sawaki & Sinharay (2013) question the reliability of overall speaking scores. The speaking section contains six individually rated tasks that are reported as a single aggregate score. Though a learner may be strong at one or two tasks but extremely weak at another task, admissions to an institution are based on overall section scores and not individual task scores.

Brown (2005) defines validity as a "degree to which a test measures what it claims, or purports, to be measuring" (Brown, 2005, p. 200). The inclusion of the TOEFL speaking section is to assess whether scores reflect students' abilities to speak in an academic context. Brown additionally mentions that "a test cannot be valid unless it is first reliable" (Brown, 2005, p. 220). Studies have shown that scores from the speaking section are valid and reliable measures of learners' abilities to interact and communicate at English speaking universities (Sawaki & Sinharay, 2013). Zareva (2006, p. 47) additionally agrees that the TOEFL test is a true

representation of test takers abilities to communicatively function in an academic setting successfully.

Analysis

The following portion of the paper explores the TOEFL iBT speaking section using Wesche's (1983) and Swain's (1984) frameworks (as cited in Bailey & Curtis, 2015). Frameworks provide a base for analysis and evaluating tests with a common set of criteria.

Wesche's Four Components

Wesche's framework is a basic guide for analyzing the structure of a test. It consists of four components: (a) stimulus material, (b) task posed to the learner, (c) learner's response and, (d) scoring material. Table 3 outlines the TOEFL iBT as they relate to these components.

Stimulus material. Bailey (2015) defines stimulus material as the "linguistic or nonlinguistic information presented to learners to get them to demonstrate the skills or knowledge we want to assess" (p. 29). The skills assessed are test takers ability to effectively communicate in every day academic life. This ranges from communication with librarians and professors to the ability to understand lectures and academic writing. The stimulus material for the speaking section is the lectures, readings and writing prompts. For the integrated speaking task, test takers listen to a lecture and read a passage about the lecture they listened to. The independent tasks are given through oral and written directions. All prompts and passages read and heard prepare test takers for their oral response.

Task posed to the learner. This is the "unobservable mental work which demonstrates a test takers skill or knowledge through successful completion of a test item or prompt" (Bailey & Curtis, 2015, p. 29). Prior to learner's responses the information they read and hear must be understood and synthesized. The integrated tasks are different than the independent tasks. There

are four integrated tasks (numbered 1 – 4 in Table 3). The first two require the learner to listen to lectures and read passages before synthesizing information. For the next two tasks the learners listen to passages before speaking about a problem/solution topic and summarizing an academic course. Finally, there are two independent tasks, which require learners to read and understand the prompt.

Learner's response. A learner's response is the observable evidence behind student learning and understanding (Bailey & Curtis, 2015, pg. 29). It is the stage of production from the previous unobservable mental work. Under strict preparation and response time restrictions learners need to respond orally in an organized and clear fashion to the prompts. Their responses are recorded through a microphone and digitally saved.

Scoring Material. This section is scored holistically on a 4-point scale. Zero is the absolute minimum score that a test taker could receive, indicating the speaker did not respond or responded in a way that was off-topic. To ensure inter-reliability, three to six human raters score the test using an independent or integrated speaking rubric. The rubrics are designed to evaluate speech delivery, language use and topic development. The raw scores for each task are summed and converted to a 0 – 30 point scale.

Swain's Four Principles

Swain's framework is comprised of four principles for designing communicative tests: (a) start from somewhere, (b) concentrate on content, (c) bias for the best and, (d) work for washback. Table 4 outlines the TOEFL iBT as they relate to these principles.

Start from somewhere. This is Swain's (1984) first principle for designing communicative tests. It was built on the idea that assessments should be founded on theoretical principles (Bailey & Curtis, 2015). In response to validity concerns, ETS changed the design of

TOEFL to include a speaking section. This section reports students' abilities to effectively communicate in an academic setting. Given test taker's future context, ETS designed a test that would assess students' ability to communicate at the collegiate level. Assessing communicative competence in a higher education setting was the base for designing the test.

Concentrate on content. Knowing test takers goals, proficiency levels and demographics is important for designing tests with relevant and reliable content. The TOEFL speaking test concentrates on assessing English in an academic environment. This is appropriate as test takers are usually individuals desiring to study in North American English speaking institutions. It assesses their communicative competence through integrated and independent tasks (listening and reading) that require them to organize their thoughts, summarize ideas, and defend opinions through speech. The tasks posed mimic real life academic scenarios that often North American students encounter on a daily basis. The content of the test captures academic life culturally and socially as students perform integrated and independent tasks.

Bias for the best. According to Swain (1984) a test should be designed to exemplify learners' best performance. To bring out learners' best performance, the TOEFL speaking section incorporates: (a) multimedia (audio/image/written) directions (b) a toolbar to track time/question number and adjusts volume, (c) permits note taking, (d) incorporates different audio accents and, (e) accounts for test takers with disabilities.

Learners who feel relaxed and at ease may perform their best on a test. TOEFL aims to reach a bias for the best by permitting students to take notes and using visual and auditory aids, however, some test takers have argued against it. One test taker I interviewed commented on her anxiety level during this section (Author's personal communication, September 23, 2014). She stated she could hear other test takers responses and they could hear hers. Additionally, she felt

that speaking to a computer did not portray her best performance because unlike humans, computers are incapable of providing feedback, such as a nod of the head or eye contact.

Work for washback. Washback is the effect, either positive or negative, on how learners study or go about learning a language and teachers teach a course. Buck (1988) defines it as a "natural tendency" for teachers to teach to the test, while Brown (2007) defines it as "the effects of an assessment on teaching and learning prior to the assessment" (p. 451). To promote positive washback, ETS produces extensive online information for learners and teachers. Their website answers frequently asked questions, includes interactive sample tests and rubrics for rating speaking tasks. Additionally, preparation courses provide books, practice tests and intensive instruction with interactive feedback to prepare students for the test.

Although ETS has gone through great measures to promote positive washback, a study conducted by Erfani (2012) found that memorized words and chunks negatively affected washback. This did not reflect learner's development or language acquisition. A study conducted by Andrew et al. (2002) similarly found that rote memorization did not produce meaningful learning (as cited in Zareva, 2006b).

Conclusion

Since the early 1960s the TOEFL has evolved into an internationally and widely used assessment tool for analyzing learners' abilities to effectively communicate in an academic environment. Through their high standards, ETS strives to maintain test reliability and validity. Swain's and Wesche's frameworks provided a solid foundation in understanding ETS' initiatives to promote positive washback, bias for the best, and reliable test results. After deep analysis of this assessment, I now better understand the anxiety that test takers experience.

References

Bailey, K., & Curtis, A. (2015). *Learning about language assessment*. (L. Le Drean, E. Henly, & R. Vanessa, Eds.) (second.). Boston, MA: Sherrise Roehr.

Brown, D. (2007). *Teaching by principles* (3rd ed.). White Plains, NY: Pearson Education.

Brown, H. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York: McGraw-Hill.

Buck, G. (1988). Testing listening comprehension in japanese university entrance examinations. *JALT*, *10*(1), 17.

Educational Testing Service. (2004). iBT/next Generation TOEFL test. Educational Testing Service.

Educational Testing Service. (2005). *Setting the final cut scores*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Tests/TOEFL/pdf/setting_final_scores.pdf

Educational Testing Service. (2008). TOEFL iBT tips: How to prepare for the TOEFL iBT . Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2010a). ETS proficiency profile proctor administrator manual. Educational Testing Service.

Educational Testing Service. (2010b). TOEFL internet-based and paper-based tests. Princeton, NJ.

Educational Testing Service. (2011). Information for score users, teachers and learners. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v5.pdf

Educational Testing Service. (2014a). 2014 -15 TOEFL iBT registration form. Educational Testing Service.

Educational Testing Service. (2014b). About the TOEFL iBT test. Retrieved September 19, 2014, from https://www.ets.org/toefl/ibt/about?WT.ac=toeflhome_ibtabout2_121127

Educational Testing Service. (2014c). Reading section. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2014d). Scores. Retrieved September 21, 2014, from https://www.ets.org/toefl/institutions/scores

Educational Testing Service. (2014e). TOEFL PBT test fees. Retrieved September 19, 2014, from https://www.ets.org/toefl/pbt/about/fees/

ETS. (2007). *TOEFL computer-based and paper based tests*. Princeton, NJ.

ETS. (2011). TOEFL program history. *TOEFL iBT Insight*, *6*(1).

Interpret scores. (2014). Retrieved September 21, 2014, from https://www.ets.org/toefl/institutions/scores/interpret/

Jamieson, J., & Poonpon, K. (2013). Developing analytic rating guides for TOEFL iBT ® integrated speaking tasks. Princeton, NJ: Educational Testing Service.

Qian, D. (2010). *English language assessment and the chinese learner*. New York: Routledge.

Sawaki, Y., & Sinharay, S. (2013). Investigating the value of section scores for the TOEFL iBT ® test. Princeton, NJ: E.

Seyed Erfani, S. (2012). A comparative washback study of IELTS and TOEFL iBT on teaching and learning activities in preparation courses in the iranian context. *English Language Teaching*, *5*(8), 185–195. doi:10.5539/elt.v5n8p185

Tannenbaum, R. J., & Wylie, E. C. (2004). *No mapping test scores onto the common european framework: Setting standards of language proficiency on the test of english as a foreign language (TOEFL), the test of spoken english (TSE), the test of written english (TWE), and the test of english for Int*. Princeton, NJ: Educational Testing Service.

Tumposky, D. (2009). Ensuring quality and reliability in scoring TOEFL IBT ® speaking and writing. Educational Testing Service.

Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL academic speaking test (TAST). Princeton, NJ: Educational Testing Service.

Zareva, A. (2006a). What is new in the new TOEFL-iBT 2006 Test Format ?, *2*(2), 45–57.

Zareva, A. (2006b). What is New in the New TOEFL-iBT 2006 Test Format ?, *2*(2), 45–57.

Table 1

General TOEFL iBT Format (Educational Testing Service, 2008)

Tasks	Description	Minutes
Reading Section		
3-4 passages, 12–14 questions each	Passage = ~700 words	60–80
Listening Section		
4 -6 lectures, 6 questions each	Lectures = 3-5 minutes each ~500 – 800 words	60-90
2-3 conversations, 5 questions each	Conversations = 3 minutes long ~20 -25 exchanges	
Speaking Section		
6 tasks, 2 independent & 4 integrated	Independent • Personal preference • Choice Integrated • Campus situation topic: Fit & explain o Read/listen/speak • Academic course topic: General o Read/listen/speak • Campus situation topic: Problem/solution o Listen/speak • Academic course topic: Summary o Listen/speak	20
Writing Section		
1 integrated & 1 independent	Integrated Task Read/listen/write (150 – 225 words) Independent Task Writing from experience Opinion on an issue (min. 300 words)	20 30

Table 2

Speaking Task Types (Educational Testing Service, 2008)

Task	Description	Seconds (Prep/Response)
Independent		
1	Describe and defend a personal choice from a given category such as important people, places, events or activities	15/45
2	Defend a personal choice between two contrasting behaviors or courses of action	
Integrated		
1	Read a passage (75-100 words) on campus related issues, listen to a passage (60-80 seconds) on the issue from the reading passage, and summarize the speakers' opinion within context of the reading passage	30/60
2	Read a passage (75-100 words) which explains the process of a term, listen to a lecture excerpt (60-90 seconds) that provides descriptive examples of the term and its process from the reading excerpt, and combine the information to convey the important information from the reading and lecture	30/60
3	Listen to a student-related problem and solutions (60-90 seconds) and express an opinion on solving the problem	20/60
4	Listen to a lecture (90-120 seconds) that explains and gives examples of a term/concept, and summarize the lecture by showing the relationship between the examples of the topic	20/60

Table 3

Wesche's Framework (1983)

Task	Integrated	Independent
Stimulus material		
(1)	1 campus reading passage & 1 campus listening passage	1 personal preference prompt (e.g. a place you enjoy)
(2)	1 academic reading passage & 1 academic listening passage	1 personal choice prompt (e.g. defend contrasting behaviors)
(3)	1 campus listening passage	
(4)	1 academic listening passage	
Task posed to the learner		
(1)	Understand written & spoken passages (30 sec). Synthesize information and respond verbally.	Understand prompt (15 sec). Respond verbally
(2)	Understand written & spoken passages (30 sec). Synthesize information and respond verbally	Understand prompt (15 sec). Respond verbally
(3)	Understand spoken passage (20 sec). Respond with an organized verbal opinion	
(4)	Understand spoken passage (20 sec). Verbally summarize the passage	
Learner's response		
(All)	60 second response	45 second response
Scoring material		
(All)	1 – 4 holistic scoring rubric. Focus: delivery, language use and topic development	1 – 4 holistic scoring rubric. Focus: delivery, language use and topic development

Table 4

Swain's Framework (1984)

Principle	In iBT Speaking Section
Start from somewhere	Communicative competence
Concentrate on Content	Academic content and context is relevant to target demographic
Bias for the best	Multimedia (audio/image/written) directions Toolbar to track time/question number and adjust volume Note taking permitted
Work for washback	Preparation courses and materials Extensive online information Interactive practice tests