EVALUATION and CRITIQUE OF TOEFL

Murat KILINC

University of Iowa
EVALUATION and CRITIQUE OF TOEFL

ABSTRACT

Students from all over the world are dealing with some special exams such as GRE, GMAT, IELTS to be eligible to receive quality education in the U.S and western countries. Test of English as Foreign Language (TOEFL) is arguably the biggest challenge for these students to overcome. Validation of a test which has an impotent place for student’s success is substantial issue. The purpose of the study is to investigate TOEFL validity and reliability by using some research conducted by specialists in English Testing Service (ETS). Though the studies need to be replicated in other conditions, all consequences reached by the researches reveal that TOEFL are valid exam in terms of criterion validity, content validity, and score reliability.

Keywords: content validity, criterion validity, score reliability, and TOEFL
Millions of people have dreams to take advantages of being student in the U.S and some other countries which are very famous for their marvelous education quality. For being a student in countries whose requirement is firstly to have sufficient English skills, students are required to take certain exams. Test of English as a Foreign Language (TOEFL) is one of them and the most popular language exam all over the world. Moreover, TOEFL is the most commonly taken international English language test worldwide (e.g. Pierce, 1992; Hamp – Lyons, 1998).

The test aims to evaluate English proficiency of people whose native language is not English. To quote the original TOEFL framework document (Jamieson, Jones, Kirsch, Mosenthal & Taylor, 2000, pp. 10-11): “The purpose of the test will be to measure the communicative language ability of people whose first language is not English.”

The test has two different formats which are paper-based test and internet based test. The sections are reading, listening, and writing for paper-based test. The internet based test includes speaking exam as well. Reading and listening part includes multiple choice questions. For speaking tasks, test takers wear headphones and speak into a microphone. Digital recorded system is used to evaluate student’s response. For writing tasks, responses are typed by students for a given record, the type responses are also sent to ETS’s online scoring network.

The reading part measures test taker’s ability to understand university-level academic texts. The test takers read 3-5 passages of approximately 700 words each and answer 12-14 questions about each passage. The listening section measures test takers’ ability to understand
spoken English in an academic setting. Test takers listen 4-6 lectures and 2-3 conversation, each about 3 minutes long. The speaking section measures test takers’ ability to speak English effectively both inside and outside of the classroom. The speaking section consists of six tasks. For two of them, test takers do not receive any oral or written test material. They respond to general questions on a familiar topic. The other tasks in the speaking section include both an oral and a written stimulus. The writing part measures test takers’ ability to write about an academic topic. The writing part consists of two tasks; integrated and independent writing tasks.

Each section in the test is scored on a scale of 0-30. The test takes 4 hours to complete. After the reading and listening sections, 10 minute break is available for test takers.

The test is used in broad ways such as admission for a university, selection for a job. Therefore, the audiences for the test are test takers, admission boards in the universities, employers and so on.

In the present paper, the validity evidence for TOEFL is investigated by the help of three research papers published by English Testing Service (ETS). Consecutively, predictive validity, content validity, and score reliability are going to be assessed.

Validity Evidence for the Test

Predictive Validity

The validity studies for the test are focused on criterion-related validity because of the fact that a test like a TOEFL is firstly used for admission. It is easy and essential for test developer to
present validity evidence for this purpose. In this framework, the first study provided by ETS that is going to be presented is criterion-related validity of the TOEFL speaking scores for international teaching assistants screening exam, which was conducted by Xaiming Xi who is a researcher at ETS.

Xaiming Xi worked with four universities for his research (University of California, University of North Caroline, Drexel University, and University of Florida). In his research, the speaking score of TOEFL was correlated with the score of an exam international student has to take before they start their education. Local international teaching assistant assessments have different names and requirements for these four universities. For example, University of California conducts an exam whose name is the Test of Oral Proficiency (TOP) which consists of three tasks which are self-introduction, short presentation for an easy topic given by test implementer, and prepared presentation for a topic test in the test taker’s field. Other universities conduct similar exams but not the same one.

As an example from research, University of California is going to be presented as steps for its exam structure were defined above and the results for other universities are almost the same. According to Xiaoming (2008), “The Wald test suggests that the TOEFL speaking score were a significant predictor of the TA assignment outcomes. A positive B coefficient (.60) suggests that the likelihood is high that a student with a high TOEFL speaking score will be in a higher TA assignment category”. The outcomes for other three universities are similar to the result of University of California.
EVALUATION and CRITIQUE OF TOEFL

From the outcomes of the research, the TOEFL speaking scores had strong relationship with international teaching screening. Therefore, it is a fact that speaking part of TOEFL is a substantial predictor of the international teaching assignment results.

Content Validity

It is hard to find a study which covers all TOEFL sections. Just as the predictive validity described above, many validity research for TOEFL focus on only one or two sections of the test. For content validity, a research focusing on content validity of writing and speaking sections of the exam is going to be showed.

A teacher-verification study of Speaking and Writing Prototype Tasks for a New TOEFL is going to be investigated. The study, purpose of which was formative, was conducted by ETS researcher in 2000 when ETS was releasing TOEFL IBT. Seven instructors who were highly experienced in ESL instruction, and had relevant educational qualifications were chosen for the study. Three questions researchers were seeking answers were related to content validity, authenticity, and educational relevance. Only the first question investigates the content validity: “Is the content of the prototype tasks being field tested for the new TOEFL perceived to assess the domain of academic English required for studies at an English-medium university in North America?” The researchers were seeking an answer for this question. They prepared a questionnaire to find an answer for the question.

In the questionnaire, the researcher asked each instructor to evaluate the effectiveness of each of the prototype task types; then to explain briefly why the instructors considered each task type effective or not; then to rate how well they thought the task represented the domain
of academic English required for studies at an English-medium university. Moreover, other questions asked to instructors were to explain why which tasks do this best, and what might be missing from the tasks. Before implementing the questionnaire, the researchers gave a self-ratings scale to the instructors attending the research. Depending on their answers on the self-ratings scale, they considered themselves as either “competent” (four people), “expert” (one person), or in between “competent” and “expert” (two people) with regards to their skills in assessing English language proficiency.

The structure of prototype tasks was the same of real TOEFL-IBT test. Two independent, four integrated speaking topics and one independent and one integrated writing topic was showed to the instructors.

The researchers analyzed the instructors’ responses in the questionnaire to answer the questions related to content validity and completely defined above. After analyzing the answers of instruction in questionnaire, researchers reached to the conclusion that all instructors attending to the research thought that the tasks represent quite well of the domain of academic English required for studies at an English-medium university. Moreover, they added additional comments and advices to the questionnaire. To illustrate, one instructor said for overall evaluation, “If a student can read/listen to a fairly complicated text and respond clearly, it seems like a good indication they’ll be able to do academic work.” Also, to give an example from suggestions, an instructor thought that there should be more choices for the examinees to write about based on their individual preferences or interest.
Taking into account researchers' discussion in the research paper, the conclusion can be drawn that ESL instructor thought the tasks in speaking and writing sections are appropriate to assess students' readiness for their undergraduate and graduate studies in terms of content of the test.

Score reliability

The quality of a test is an important issue to determine how the reliability of test scores is. In essence, “the concern of reliability is to quantify the precision of test scores and other measurements” (Haertel, 2006, p. 65). Standard error of measurement expresses precision of test scores.

A study provided by ETS gives us information about reliability of estimate. Mary Enright and Eileen Tyson stated the following:

The reliability estimation for the Reading and Listening sections that contain selected response questions was carried out using a method based on item response theory (IRT) (Lord, 1980). For the Speaking and Writing sections that contain constructed response tasks, generalizability theory is used (Brennan, 1983). (p. 4)

The reliability of estimates based on an operational data from 2007 was .85 for reading, .85 for listening, .88 for speaking, and .74 for writing. Standard error of measurement (SEM) for these subtests of TOEFL are 3.35 for reading, 3.20 for listening, 1.62 for speaking and 2.76 for writing. Moreover, the reliability of estimates for total score of entire test is .94 and standard error of measurement is 5.64. When this result is analyzed, it can be said that the best information is gained from the total score because it reflects four language skills and it is more reliable.
With this information, the same study mentions about alternate form reliability. To
determine this reliability, it is not calculated based on examinees’ score on different forms of a
test. Instead, researcher analyzed examinees taking two tests of the same form in a short period of time. According to Zhang (2008), an analyse of the scores of these repeat test takers on the two test forms provides an approximation of alternate form reliability. Zhang (February 2008) compared the test scores of more than 12,000 examinees who were identified as having taken two TOEFL iBT tests within a period of one month. The correlations of their scores on the two test forms were 0.77 for the listening and writing sections, 0.78 for reading, 0.84 for speaking, and 0.91 for the total test score. Even these results are lower than internal consistency measure, they still indicate a high degree of consistency.

Depending on information gained from the research, a conclusion can be drawn that the reliability of the test score is existant for the test.

Discussion

In this study, we have three different validity evidences from the developer perspective: predictive validity, content validity, and score reliability. Firstly, content and test reliability provided by ETS is going to be discussed.

Content validation has been done by using ESL instructor. We should take into account following points for the selection of instructors used in the study. The researcher gives information about the instructors chosen for the study. Form this information, it would be convenient and more revealing to ask the following questions: “If we used different instructors for this study, could we reach the same or similar consequence for content validation?”,”if
more experienced instruction were chosen, would the result be the same?” and “if a large sample of instructors were used instead of using seven ESL instructor, would the result be the same?” Many questions like those may come to mind when selecting instructors used in the research.

Internal consistency and alternate reliability were indicated. This study shows that internal consistency and score reliability are existent for TOEFL. However, Robert M. Thorndike and Tracy Thorndike-Christ (2010) states, “Alternate form of a test should be thought of as forms built according to the same specifications but composed of separate samples from the defined behavior domain” (p. 124). Hence, alternate reliability coefficient could be supported with a real alternate form test instead of using the scores of test takers who have taken the exam in a period of one month. Many different English testing exams such as International English Language Testing Systems (IELTS), that is a good alternate form test of TOEFL and is second popular English language exam all over the world, can be used. Moreover, many students who took both exams in a short period of time can be found. Hence, ETS would conduct a study with the providers of IELTS. This will almost give exact information about alternate reliability coefficient instead of an approximate one.

Predictive validity was provided; the researchers used speaking score of TOEFL as a predictive of an English exam which is teaching assistants screening exam. The study reveals that the score of speaking part is good predictor of exam score conducted by four universities. The research mentioned above provided correlation coefficient between TOEFL subtests and English exams conducted by universities.
Conclusion

All in all, taking into account all things mentioned above, the studies inspected in the paper and overwhelming majority of similar studies supports the claim that TOEFL is the most widely respected English-language test in the world.
REFERENCES


