Test review: Test of English as a Foreign LanguageTM: Internet-based Test (TOEFL® iBT®)
J. Charles Alderson
Language Testing 2009; 26; 621
DOI: 10.1177/0265532209346371

The online version of this article can be found at:
http://ltj.sagepub.com

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for Language Testing can be found at:
Email Alerts: http://ltj.sagepub.com/cgi/alerts
Subscriptions: http://ltj.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.co.uk/journalsPermissions.nav
Citations http://ltj.sagepub.com/cgi/content/refs/26/4/621
Test review

Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®)

J. Charles Alderson  Lancaster University, UK

TOEFL iBT: TOEFL iBT is the latest version of the TOEFL, whose history stretches back to 1961. The first paper-based TOEFL was administered in 1964, the related Test of Spoken English (TSE) was developed in the late 1970s,1 and the Test of Written English (TWE) was made available as an optional extra to the TOEFL from 1986.2 A computer-based version of TOEFL (CBT) was introduced in 1998, and the final administration of CBT was in September 2006. The TOEFL iBT was introduced in the USA, Canada, France, Germany and Italy in late 2005, followed by a gradual roll-out across the world. A detailed history is available in Chapelle et al. (2008, pp. 359–361). Currently the TOEFL test is offered in two testing formats: Internet-based testing (iBT) and paper-based testing (PBT). The format taken depends on the location of the test centre.

Test purpose: The TOEFL iBT measures the ability of non-native speakers of English to use and understand English as it is spoken, written and heard in academic settings.

Use: TOEFL scores are accepted by more than 6000 colleges, universities, licensing agencies and immigration authorities in 136 countries.

1 The TSE will continue to be available separately as long as the paper-based version of TOEFL is available, but TOEFL iBT now incorporates a test of speaking ability.
2 The TWE became a mandatory part of the TOEFL paper-based test in 1998 with the introduction of the TOEFL computer-based test, which included a writing task. The TOEFL iBT includes two writing tasks.

© The Author(s), 2009. Reprints and Permissions: http://www.sagepub.co.uk/JournalsPermissions.nav
DOI:10.1177/0265532209346371
**Length and administration:** Test takers have up to four hours to complete the TOEFL iBT. The test is administered 30–40 times a year at more than 4300 test centres worldwide.

**Registration:** Online registration is available at www.ets.org/toefl/ibt/register.html, and this site also contains information on test locations and dates, rescheduling or cancelling registration, disability accommodations, as well as test preparation and FAQs. See also ETS (2008a).

**Scores:** Scores are reported on a scale of 0–30 for Reading, Listening, Speaking and Writing. A total score out of 120 is also reported. ETS offers a TOEFL iBT standard-setting CD-ROM to assist institutions with setting cut scores, and the TOEFL Score Comparison Tables are also available, showing comparisons of TOEFL iBT, TOEFL CBT, and TOEFL PBT scores. Scores are reported online, as well as in paper and electronic form, and can be accessed by test takers online 15 business days after the test.

**Author and publisher:** Educational Testing Service (ETS), Princeton, NJ, USA. Mailing address: TOEFL® Services, Educational Testing Service, P.O. Box 6151, Princeton, NJ 08541–6151, USA. Fax: +1–610–290–8972. Phone: +1–609–771–7100 or +1–877–863–3546, Monday–Friday, 8 a.m.–8 p.m., Eastern Time (New York), or +1–609–771–7714 (TTY), 24 hours a day, 7 days a week.

**Price:** The price of TOEFL iBT varies by location, ranging at the time of writing from $150 in the USA, Central and South America, Africa, the Middle East and Asia to $185 in most of Europe, Canada, Australia and New Zealand.

**Information available:** The TOEFL website (www.toefl.org/) contains much information about the test for test takers, academic institutions, and English language teachers, as well as details of TOEFL research publications, including the Framework for Recent TOEFL Research, the TOEFL iBT Research series, TOEFL Research Reports, Technical Reports, monographs and more.

**General description**

In addition to the general purpose described above, taken from TOEFL publications, Sawaki et al. (2009, p. 5) say that the test battery
is ‘designed primarily for admission of non-native speakers of English to higher education institutions in North America’. The aim was to ‘reflect current theories of communicative language use in an academic context’ (Taylor & Angelis, 2008, p. 41). Crucially, TOEFL iBT is in many ways radically different from earlier versions of TOEFL, in that it is said to be ‘better aligned to the variety of language use tasks that examinees are expected to encounter in everyday academic life’ (Sawaki et al., 2009, p. 5). It includes mandatory speaking and writing sections, and integrated tasks that require test takers to read a text, listen to a lecture or conversation and write or speak a response to what has been read or heard. The separate test of Structure and Written Expression is no longer given, and test takers read or listen to longer written and spoken texts in the reading and listening sections and are allowed to take and refer to notes. TOEFL iBT is delivered via an Internet-based delivery system in ETS-certified test centres (unlike TOEFL CBT, which was delivered on dedicated computers in specially equipped test centres) and candidates use the keyboard to input their answers, including to writing tasks, and a microphone to record their responses to speaking tasks. The test is linear, not computer-adaptive, and the listening and reading sections have both a long version and a short version. The long version includes experimental, pre-test, research, or equating questions. Each test taker receives a long version of either listening or reading, but not both, thus ensuring that all test takers are given tests that are comparable in length. As it is impossible to tell which questions are experimental, candidates have to do their best on all.

Reading

The reading section has fewer but longer texts than previous versions of the TOEFL (600–700 vs. 300–400 words), taken from university-level textbooks that introduce a discipline or topic, and examinees are allowed between 60 and 100 minutes for this section. Each test has 3–5 texts on different academic topics, classified by author purpose: exposition, argumentation, and historical narrative. There are between 12 and 14 questions per text, designed to measure basic comprehension, inferencing, and reading to learn. There is an

---

3 TOEFL scores are accepted by many English-medium institutions of higher education throughout the world.
average number of 39 items for the short form of the iBT. There are three item types: traditional single-answer four-option multiple choice; single-answer four-choice ‘Insert a sentence where it best fits in the text’; and partial-credit reading-to-learn items with more than four choices and more than one possible correct answer.

Reading to learn is assessed by prose summary items (‘An introductory sentence for a brief summary of the passage is provided below. Complete the summary by selecting the three answer choices that express important ideas in the passage.’); and by Schematic Table items (‘Complete the table below to summarize information about the two types of art discussed in the passage. Match the appropriate statements to the types of art with which they are associated. Drag your answer choices to the spaces where they belong.’).

**Listening**

The listening section includes 4–6 lectures, some with classroom discussion, each 3–5 minutes long and containing 500–800 words. There are 6 questions per lecture. In addition there are 2–3 conversations, each about 3 minutes long, consisting of 12–25 exchanges, and with 5 questions per conversation. Item types include traditional single-answer four-option multiple-choice; multiple-choice with more than one correct answer; sequencing items which require the ordering of events or steps in a process; and items that require the matching of objects or texts to categories in a chart.

Skills tested include listening for basic comprehension (main idea and important points), listening for pragmatic understanding (speaker’s attitude, degree of certainty, purpose); and connecting and synthesizing information (recognizing organization, understanding relationships, making inferences and drawing conclusions).

Pictures on the screen help test takers to visualize the setting and roles of the speakers. Most passages are played only once, and questions are not presented in advance but note-taking is allowed throughout. The time allowed for this section is between 40 and 60 minutes.

**Speaking**

The speaking section contains six tasks, two of which are ‘independent’, that is, candidates express an opinion on a familiar topic; and four are integrated tasks, where candidates speak based upon something heard or read. The speaking section measures the abilities to speak
about everyday familiar topics, and to summarize, evaluate, compare and synthesize information from multiple sources in a spontaneous manner. Reading passages are 75–100 words long; listening passages take between 60 and 120 seconds. Candidates’ comprehension is not separately assessed in these texts. Candidates are given up to 30 seconds to prepare their response and up to one minute to respond. Each task is rated on a 0–4 point scale, and scored on four criteria: general (including intelligibility, task fulfilment and coherence); delivery (clarity, fluency, pronunciation, intonation, stress); language use (range and control of grammar and vocabulary); and topic development (relationship and progression of ideas, relevant content). The speaking section takes 20 minutes.

Writing

The writing section involves two tasks, one an independent task where candidates write an essay that expresses opinions or choices, which should be supported and explained. Candidates are given 30 minutes to prepare and write their response.

The second task is an integrated task, where candidates read a short text of around 230–300 words (they are allowed 3 minutes for reading) on an academic topic. The text then disappears from the screen and they listen to a speaker on the same topic from a different perspective. The listening passage is about 230–300 words long and takes about 2 minutes. Note-taking is allowed throughout, but once the listening passage has finished, the written text reappears and candidates write a summary of the main points made in the listening passage and how these relate to the key points in the written text. Suggested response length is 150–225 words and candidates have 20 minutes to respond. Answers must be typed.

Responses are rated on an overall, holistic 0–5 scale, with different descriptors for the independent and the integrated tasks. The independent essay is scored on the overall quality of the writing: development, organization and appropriate and accurate use of grammar and vocabulary. Responses to the integrated task are also scored on the quality of the writing, but, additionally, for the completeness and accuracy of the content. E-rater® will be used as one of the two raters for the TOEFL iBT independent writing task starting in summer 2009. The integrated task will continue to be scored by two humans, and so a test taker’s writing score will be based on one automated rating and three human ratings. E-rater uses natural
language processing methods to evaluate the linguistic features of an essay: grammar, usage, mechanics, vocabulary, organization and development (see Enright & Quinlan, 2009).

**Appraisal**

The TOEFL is a heavily researched test, with roughly 100 research and technical reports on earlier versions of the TOEFL, and a monograph series, starting in 1997, that reported on developments, *inter alia*, in the TOEFL 2000 project. Clearly a brief review can barely mention, let alone do justice to, such a wealth of research.

The TOEFL 2000 Project itself was complex, and adjusted its name and direction several times. From TOEFL 2000 came, in 1998, TOEFL CBT (barely distinct in construct from the paper-based TOEFL), then Next Generation TOEFL, LanguEdge test preparation software (a prototype of TOEFL iBT) and finally TOEFL iBT. Hopefully, this proliferation of names will now settle down, even if TOEFL iBT itself evolves in line with ongoing research into its qualities. Although this review concentrates on TOEFL iBT, inevitably reference will need to be made to research into earlier versions, especially research based on the prototype LanguEdge test preparation software, as research is only now being published in the new TOEFL iBT Research Report series.

Chapelle et al. (2008) describes the issues discussed during the Project, the research carried out, and the ways in which the final product – TOEFL iBT – was shaped by that research process. The book is a major contribution to the literature and an excellent record of the process of and changes in the development of the iBT and the building of a validity argument based on the Toulmin framework (Toulmin, 1958).

ETS completed a questionnaire for this reviewer, based on the Guidelines for Good Practice of the European Association for Language Testing and Assessment (EALTA, 2006). The responses were full, indicating compliance with good practice.

**Validity and reliability**

ETS (2008b) presents the validity argument for the TOEFL iBT test by ‘stating the propositions that underlie the proposed score
interpretation and uses, and by summarizing the evidence related to each proposition’ (2008b, p. 1). These propositions include the relevance and representativeness of test content; the appropriateness of task design and scoring rubrics; the relationship to academic language proficiency of the linguistic knowledge, processes and strategies that test takers use to respond to test tasks; the relationship between the test structure and theoretical views of the relationships among English language skills; and the relationship between TOEFL iBT scores and other criteria of language proficiency.

Chapelle et al. (2008) makes a case for the proposed interpretation and use of TOEFL iBT test scores, and readers of this review are strongly recommended to study that volume. The numerous framework papers cited in Chapelle et al. and ETS (2008b) describe the development of a rather radically different set of constructs from TOEFL PBT, and a much more task-centred approach to test design. TOEFL iBT score interpretation was to be underlain by a theoretical construct, based on a theory of communicative competence, ‘which accounts for language performance across a wide range of contexts, includes complex abilities responsible for a particular range of goals and takes into account relevant contexts’ (Jamieson et al., 2008, p. 57).

On the whole, TOEFL iBT has achieved this, with a clearer focus on the academic environment, based on research into the language of academic tasks, careful prototyping, trialling, revisions, more trialling, and so on. The inclusion of a compulsory speaking section, the integration of skills in numerous tasks, the use of longer written and spoken texts which are more obviously authentic and academic in nature, and have less obvious bias towards a North American setting, a radical reduction of focus on grammar: all these innovations and more are welcome. Importantly, this increased emphasis on the academic context will hopefully reduce the misuse of TOEFL in non-academic testing contexts like immigration.

ETS (2008c) argues that TOEFL iBT scores are reliable and comparable because standardized administration and security measures are adhered to; detailed test specifications are used to guide test development; score reliability and generalizability are monitored (and reported in various studies); appropriate scales are developed for reporting scores; score comparability across test forms is maintained by equating of the reading and listening sections and by statistical analysis of both tasks and raters on the speaking and writing sections.
Reported IRT-based reliability indices for Reading and Listening were acceptably high at 0.85 in both cases, based on operational data from 2007 (ETS, 2008c). The generalizability theory-based reliability for Speaking was 0.88, but weaker for Writing (0.74), although it must be remembered that there are only two writing tasks. Test-retest reliability, based on an analysis of candidates repeating the test once within 30 days (Zhang, 2008), was moderate to high (Reading 0.78, Listening 0.77, Speaking 0.84, Writing 0.77, Total 0.91). There were small mean score increases on the second test, but the effect sizes were low (ES = 0.17 Reading, 0.12 Listening, 0.14 Speaking, 0.17 Writing, 0.17 Total).

Inevitably, since testing necessarily involves compromise, there is evidence of construct under-representation and construct-irrelevant variance. Only monologic speaking is tested, corresponding to the ability to make presentations in class and to answer personal questions or questions about lecture or reading materials. However, an ability to engage in interactive speaking is also necessary in academic study, in order to clarify one’s understanding, question what has been heard, challenge others’ ideas, and so on. In the Writing section, inevitably, the tasks are severely restricted in time, suggesting that what is assessed is the ability to create a first draft only, and not to revise, refine and polish drafts.

In terms of construct irrelevance, the test is still heavily dependent on multiple-choice item types. Cohen & Upton (2006) investigated through verbal protocols the strategies used by candidates responding to reading tasks. Unfortunately, of the 28 reading strategies predicted to occur, only three were used frequently, two occasionally, and the remainder not at all, whereas of the 28 ‘test management strategies’ 20 were used frequently, and most of these related to coping with multiple-choice items.

In the Listening section, questions are not presented before the listening passages, and so test takers have no purpose for listening other than to try to understand and remember everything. Although note-taking is allowed, it is impossible to take notes on everything, given the length and detail of the passages, and so the test may involve construct-irrelevant factors like memory and the ability to figure out the correct answer from the internal structure of (especially) the multiple-choice items.

However, one possible factor of construct irrelevance, namely computer familiarity and keyboarding skills, seems no longer to be an issue,
at least according to Wall & Horák (2008b), possibly due to the 8-year period during which the CBT TOEFL was administered, and the development of tutorials to accompany the test and test preparation courses.

Washback and impact

One of the key aims of the Next Generation TOEFL Project was to improve the impact of TOEFL on teaching. Each of the initial framework documents says that the new examination should have more positive impact on the classroom: ‘courses will more closely resemble communicatively oriented academic English courses’ (Bejar et al., 2000, p. 36). Wall & Horák (2006, 2008a, 2008b) report on a five-year study into the impact of the new TOEFL test on teaching practice in selected countries in Europe, and conclude that TOEFL iBT has indeed had the desired effects on the content of TOEFL preparation classes, in that much more emphasis was now placed on the teaching of speaking abilities, there was an increase in the attention paid to writing and on integrated skills work, and there was much less evidence of the teaching of grammar and vocabulary.

However, confirming the results of other washback studies, including the only previous study of the paper-based TOEFL (Alderson & Hamp-Lyons, 1996), Wall and Horák (2008b) found that the impact of the test itself was much less direct, and was clearly mediated by coursebooks. Moreover, the three teachers studied differed in the amount of change they showed in their methodology.

One teacher displayed very little change, continuing to offer classes where students mainly completed practice tests at the computer. Another teacher’s classes were now more interactive, especially in the discussions related to writing and speaking, but most of the interaction was of the teacher-student variety. The third teacher’s classes had changed dramatically from teacher-centred sessions focussing mainly on grammar to sessions displaying student-student interaction and other communicative characteristics.’ (Wall & Horák, 2008b, pp. vi–vii)

Importantly, ETS has made serious efforts to promote positive washback through the website, through innovative publications like ETS (2007), and through collaboration with publishers. Wall & Horák’s impact study illustrates in some detail the practices of teachers a year after the launch of TOEFL iBT in Europe. This should prove a valuable comparator and benchmark for future washback studies, such as those that ETS is currently encouraging in its announcement of the TOEFL COE 2010 research programme.
Conclusion

ETS deserves to be congratulated on the development of the TOEFL iBT. It represents a major step forward in testing language proficiency for academic purposes. Challenges remain, inevitably, in terms of the low reliability of the Writing test (the addition of e-rater® may enhance reliability), possible construct-irrelevant variance in the Reading and Listening tests, and construct under-representation in the Speaking and Writing tests, but a continuing programme of research will hopefully continue to refine and improve the test further.

References


