Review of the Test of English as a Foreign Language Internet Based Test (TOEFL iBT) writing test

Chelsea Bernal
University of Melbourne

Overview

Development background

The Test of English as a Foreign Language Internet Based Test (TOEFL iBT®) is the third version of the test, and one of two test versions which are currently available, the other being TOEFL PBT (Paper Based Test) which is being phased out and is only administered in countries without viable Internet testing. The TOEFL CBT (Computer-Based Test) was offered from 1998–2006 and was transitioned worldwide to the TOEFL iBT over its final year of administration. The primary changes from the PBT and CBT to the iBT occurred in the addition of a Speaking component and a revision of the Writing component, the latter of which is the focus of this review.The main emphasis of these changes was to make the test more representative of an academic environment, ensuring that the test is testing the same abilities that will be required in the academic courses that accept it. These new test formats consist of integrated tasks where candidates are required to read and/or listen and then speak or write in response to a question.

Test purpose

To provide a test of academic English for universities and colleges to determine language fitness for courses of study. It is also marketed as an entry/exit test for English study programs, for scholarships and professional certification, for English-language learners to track their progress and for study and work visa applicants.

Length and administration

The test is administered in secure testing locations via computer over the Internet at more than 4,500 locations worldwide. The TOEFL iBT can take up to 4 hours with a mandatory 10 minute break. It consists of 4 sections: Reading (60–80 minutes; 36–56 questions), Listening (60–90 minutes; 34–51 questions), Speaking (20 minutes; 6 tasks) and Writing (50 minutes; 2 tasks).

Author and publisher

Educational Testing Service (ETS), USA.

Information available

The ETS website (http://www.ets.org/toefl/ibt/about) contains pages for test takers, institutions and English programs, providing a range of resources, such as a downloadable bulletin for the test, speaking and writing test scoring rubrics, sample responses and research articles.

Price

Test registration fees vary by country, ranging from USD $160–$250.

Description

The TOEFL iBT writing test consists of two parts, takes 50 minutes (20 minutes for task 1 and 30 minutes for task 2) and is the final module of the test. The two tasks represent the kinds of writing assignments that students are likely to encounter in university or college (Enright & Quinlan, 2010). The first task is an *integrated* test design and the second is an *independent* essay response. Each task requires one question to be answered (no options within the tasks). Both are rated on a scale of 0–5 with different rating criteria, these scores are combined and scaled up to a total of 30. Rating is conducted using both human raters and an automated scoring technology, *eRater®*.

Task 1: The Reading-Listening-Writing Integrated Task

This is an innovative design in high-stakes testing, and attempts to simulate the academic environment that the scores will be generalised to. Before receiving the task 1 question, candidates are introduced to two different inputs over three stages (ETS, 2005a). For all three stages, candidates are encouraged to take notes. First they are shown a passage on the computer screen; this image remains on the screen for 3 minutes. The candidate then listens to part of a lecture on the topic they have just read about, while looking at the image of a professor. At the end of this lecture the original reading passage reappears along with a task rubric and question. The passage remains on-screen for the duration of the task and it is from this point that the 20 minutes begin. Candidates are informed that their response will be judged on both the writing quality and how well they related the relationship between the reading and the lecture. Guidelines, rather than mandates, for word limits are given: 150–225 words.

Task 2: The Independent Writing Task

This task requires an opinion to be expressed and supported on a given topic (ETS, 2005a). Candidates are advised that they have 30 minutes to plan, write and revise the essay, with a guideline of 300 words given. Interestingly, candidates are given less time to write more words than in a comparable task in other high-stakes tests. For example, in the essay writing task (Task 2) of the International English Language Testing System (IELTS), candidates are given 40 minutes to write a minimum of 250 words (UCLES, 2013), while in the University of Cambridge Local Examinations Syndicate's (UCLES) First Certificate in English and Certificate in Advanced English, allowances of 40 minutes for 120-180 words (UCLES, 2012a, p. 18) and 45 minutes for 220–260 words (UCLES, 2012b, p. 22) are given for the essay task. However, Hale (1992) investigated writing performance under two time conditions, 30 and 45 minutes, for the Test of Written English (TWE), the precursor to the TOEFL iBT Independent Writing Task, and found that although candidates scored higher on the second condition, their rank order for the two conditions remained relatively the same. This finding is supported by Powers and Fowles (1997) who found that a time increase from 40 minutes to 60 minutes for the Graduate Record Examination (GRE) essay task made no difference to the meaning of the score, and also by Elder, Knoch and Zhang (2009) who have shown, at least in a diagnostic English language assessment environment, that the amount of time allocated for a writing test does not affect the validity and reliability of a candidate's score (when the time given is reduced from 55 minutes to 30 minutes in a 300-word essay task).

Rating

The two TOEFL iBT writing test tasks are rated twice independently. The public version of both the integrated and the independent writing rubrics (ETS, 2004; see Appendix I and II) depict a 6-point holistic scale where 0 is assigned for responses that are completely off-topic, copied, blank or not in English. Both scoring rubrics show clear distinctions between levels and also solid exceptions and reasoning for awarding penalties.

The integrated writing rubric is primarily characterised by the ability to incorporate accurate information from the lecture and relate this to the reading. As you descend through the levels, the amount of lecture-specific accuracy required also decreases. A key distinguisher at levels 1, 2 and 3 are clear ceiling indicators, which would presumably prevent a response from being given a higher score. The indicators relate to lexis, syntax, omission or misrepresentation of key points. Distinctions are made between all levels in the areas of selecting the information presented, the coherence and precision of the presentation of the information, and in the syntactic and lexical error density.

The independent writing rubric contains positive statements referring to accomplishments at the higher levels of 4 and 5, positive and negative at level 3 and negative statements referring to weaknesses and flaws at levels 1 and 2. All levels have standards related to task response, argument development, essay structure, syntax and lexis, but these may be grouped together or separated out depending on the level. Syntax and lexis for instance are combined into one standard in levels 1, 4 and 5, separated into one standard for range and another for accuracy at level 3 and broken down into syntax and lexis at level 2. Each level shows a clear characterisation, which distinguishes it from the levels on either side of it.

The scoring of the TOEFL iBT writing tasks is conducted online by trained raters working within a centralised, internet-based scoring network which enables real time monitoring of rater performance by trained scoring leaders. Since November 2010, the tasks have been marked by both a human rater and an automated scoring technology developed by ETS, *eRater* (Haberman, 2011). The score awarded for each task is a combination of these two ratings. For the integrated task, the human rating makes the greater contribution to the final score.

Where does TOEFL iBT sit in the CEFR?

In 2008 standard setting was performed on the TOEFL iBT to map minimum cut scores from the TOEFL iBT onto the Common European Framework of Reference (CEFR). Using four rounds of cross-panel mean judgements, the cut scores in Table 1 were established (Tannenbaum & Wylie, 2008); you can see the score equivalences for both the individual writing test (out of a total of 30), and for the overall score (out of 120).

Table 1.TOEFL iBT cut scores on the CEFR Levels

CEFR Levels	Writing Test Scores	Overall TOEFL Score
A1	0-10	0-56
A2	11-16	
B1	17-20	57-86
B2	21-27	87-109
C1	28-30	110-120
C2		

How reliable are the TOEFL iBT Writing scores?

Field tests on the TOEFL iBT test indicate that writing scores have the most variance in terms of the difference between the score received by the candidate, and their actual ability. In their standard setting notebook, ETS (2005b) describes the difference between a candidate's "true" ability, versus that elicited by the test, stating that there is a 68% likelihood that a "true" score will be 3.06 points higher or lower than that received (1 standard error of measurement – SEM), and at 95% confidence, the score varies by 6.12 points (2 SEMs) on either side. This means that a candidate's writing could be rated as being Fair (17-23) with a 95% likelihood that their "true" score lies anywhere between Limited (0–16) and Good (24–30).

Strengths and weaknesses

Construct-specific

In the design phase the primary concern was that the tasks must represent writing which is integral to university or college contexts (Cumming, Kantor, Powers, Santos & Taylor, 2000). This has been borne out in the development of the integrated writing task, which has clearly been modelled on an academic classroom environment. And although a university student is rarely called on to write a 300-word essay in 30 minutes without reference to other sources (Weigle, 2002), it has been argued that this form of writing does allow the candidate to show language, structure and reasoning abilities that would be required in an educational setting (Enright & Quinlan, 2010). Although concern may be raised at what is really being tested in an integrated task, Sawaki, Stricker and Oranje (2009) investigated this and found that the TOEFL iBT integrated task defines the construct well and is not influenced by other factors.

Score reliability

Although the marking rubric clearly delineates discrete levels, which differentiate candidates, from a stakeholder perspective, this does not necessarily translate into a reliable evaluation of a candidate's actual writing ability. The variability of 12.24 points at 95% confidence (2 SEMs) when related to the writing test score is a serious limitation in its claim for reliability, especially when the scores for the other three modules have a variability of 9.16-9.2 (2 SEMs). In the standard setting facilitator notebook, ETS (2005b) states that extraneous factors, such as fatigue, the content of the test tasks, or the testing environment could be the cause, but perhaps further research and testing should be conducted to determine why writing is affected more by these factors than the other three modules. This has already begun, at least in the area of task content, with Cho, Rijmen and Novák's (2013) research into the influence of the task prompt on writing performance. However, an investigation into factors such as the testing order (which places writing at the end), the inclusion of more breaks (to determine if it is purely an issue of fatigue), and a revisit of the marking rubric is certainly warranted. If the rubric were scaled on a larger number of discrete levels, for instance, this may allow for less variation. Or it could be that a change needs to occur in the type of rating scale used, a move from a holistic scale to an analytic one, allowing for a more complete understanding of a candidate's writing ability. However, a review of the most recent research publications on the ETS site shows that the main focus areas, at least in terms of the writing test, are: *eRater* and research into automated scoring (e.g. Attali, Lewis & Steier, 2012; Enright & Quinlan, 2010; Monaghan & Bridgeman, 2005; Weigle, 2010), construct and score validity (e.g. Biber & Gray, 2013; Plakans & Gebril, 2013), and task characteristics (e.g. Cho, Rijmen & Novák, 2013).

Computer and Internet based

There are many advantages of a computer based test, such as increased security of live materials and easy administration, but the primary benefit is that you don't need to use handwriting recognition software for response analysis. Responses are typed directly into the computer, becoming analysable, either for scoring or to detect plagiarism. Many arguments have been made on the cost and time efficiency (Enright & Quinlan, 2010; Weigle, 2010) and greater accuracy (Attali, Lewis & Steier, 2012) of technology to rate writing and speaking and this is clearly an ETS priority. It is good to see, then, that appropriate steps are being taken to truly refine, test and develop the *eRater* system through years of research, rather than compromising the integrity of the testing system by introducing a fully automated system too early.

However, computer-testing brings other issues of contention: Internet speed, candidate age, keyboard knowledge and individual typing speed. According to Weigle (2002), in the TOEFL CBT, candidates were allowed to respond either by computer or by hand, but no reference is made to this in the 2013–2014 Bulletin (ETS, 2013).

Summary

The TOEFL iBT Writing test is clearly an exceptional addition to, and revision of, the original TOEFL test. The addition of the integrated task produces scores which have the potential to more accurately reflect the needs of its intended users, though a research focus on lowering the variability between actual and awarded writing test scores would certainly benefit both test takers and stakeholders. It is undeniably an incredibly well-supported test, both in terms of the materials and information provided on the website, and in regard to the research and development teams at ETS, although this is certainly hoped for, and expected, in a high-stakes test of its worth.

References

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*(1), 125–141. doi:10.1177/0265532212452396.

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT® Test: A lexico-grammatical analysis* (TOEFL iBT Report No. iBT-19). Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RR-13-04.pdf.

Cho, Y., Rijmen, F., & Novak, I. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT™ integrated writing tasks. *Language Testing,30*(4), 513–534. doi:http://dx.doi.org/10.1177/0265532213478796.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph Series RM-00-05, TOEFL-MS-18). ETS. Retrieved from http://www.ets.org/research/policy_research_reports/publications/report/2000/icix.

Elder, C., Knoch, U., & Zhang, R. (2009). Diagnosing the support needs of second language writers: Does the time allowance matter? *TESOL Quarterly, 43*(2), 351–360. doi:10.1002/j.1545-7249.2009.tb00178.x.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing, 27*(3), 317–334. doi:10.1177/0265532210363144

ETS. (2004). *iBT/Next generation TOEFL test integrated/independent writing rubrics (scoring standards).* Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/Writing_Rubrics.pdf.

ETS. (2005a). *TOEFL® iBT writing sample responses*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/ibt_writing_sample_responses.pdf.

ETS. (2005b). *TOEFL iBT standard setting facilitator notebook*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/setting_final_scores.pdf.

ETS. (2013). *2013–2014 Information and registration bulletin TOEFL iBT test.* Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_bulletin_2013-14.pdf.

Haberman, S. (2011). *Use of e-rater® in scoring of the TOEFL iBT® writing test.* (ETS Research Report ETS RR-11-25). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-11-25.pdf.

Hale, G. (1992). *Effects of amount of time allocated on the test of written English.* (Research Report No. 92–27). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-92-27.pdf.

Monaghan, W., & Bridgeman, B. (2005). *E-rater® as a quality control on human scores* (R&D Connections). ETS. Retrieved from http://www.ets.org/research/policy_research_reports/publications/periodical/2005/cwyf.

Powers, D. E., & Fowles, M. E. (1997). *Effects of applying different time limits to a proposed GRE writing test.* (GRE Board Research Report No. 93-26cR: ETS Research Report 96–28) Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/GREB-93-26CR.pdf.

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*(3), 217–230. doi:http://dx.doi.org/10.1016/j.jslw.2013.02.003.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL internet-based test. *Language Testing, 26*(1), 005–030. doi:10.1177/0265532208097335.

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English-language test scores onto the common European framework of reference: An application of standard-setting methodology* (TOEFL iBT Research Report TOEFL iBT-06). ETS. Retrieved from http://www.ets.org/research/policy_research_reports/publications/report/2008/hspt.

UCLES (2012a). *Cambridge English advanced: Certificate in advanced English (CAE) CEFR level C1: Handbook for teachers.* University of Cambridge ESOL Examinations. Cambridge, UK. Retrieved from https://www.teachers.cambridgeesol.org/ts/digitalAssets/117408_CambridgeEnglish_Advanced_CAE_Handbook.pdf

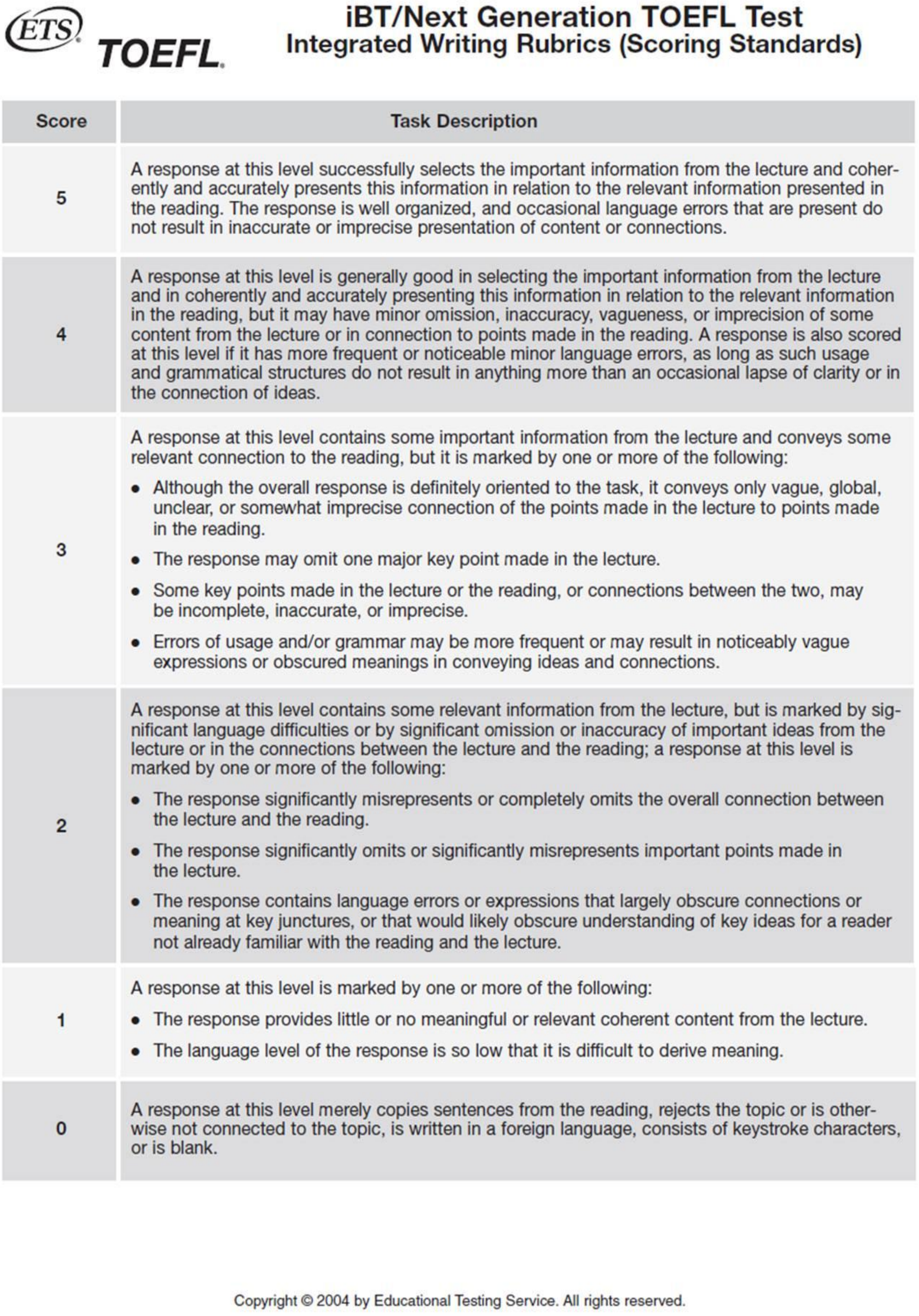
UCLES (2012b). *Cambridge English first: First certificate in English (FCE) CEFR level B2: Handbook for teachers.* University of Cambridge ESOL Examinations. Cambridge, UK. Retrieved from https://www.teachers.cambridgeesol.org/ts/digitalAssets/117578_Cambridge_English_First_FCE_Handbook.pdf

UCLES (2013). *IELTS: Information for candidates.* Cambridge, UK. Retrieved from http://www.ielts.org/pdf/Information%20for%20Candidates_2013.pdf

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335–353. doi:10.1177/0265532210364406.

Appendix I: TOEFL iBT Integrated writing rubrics

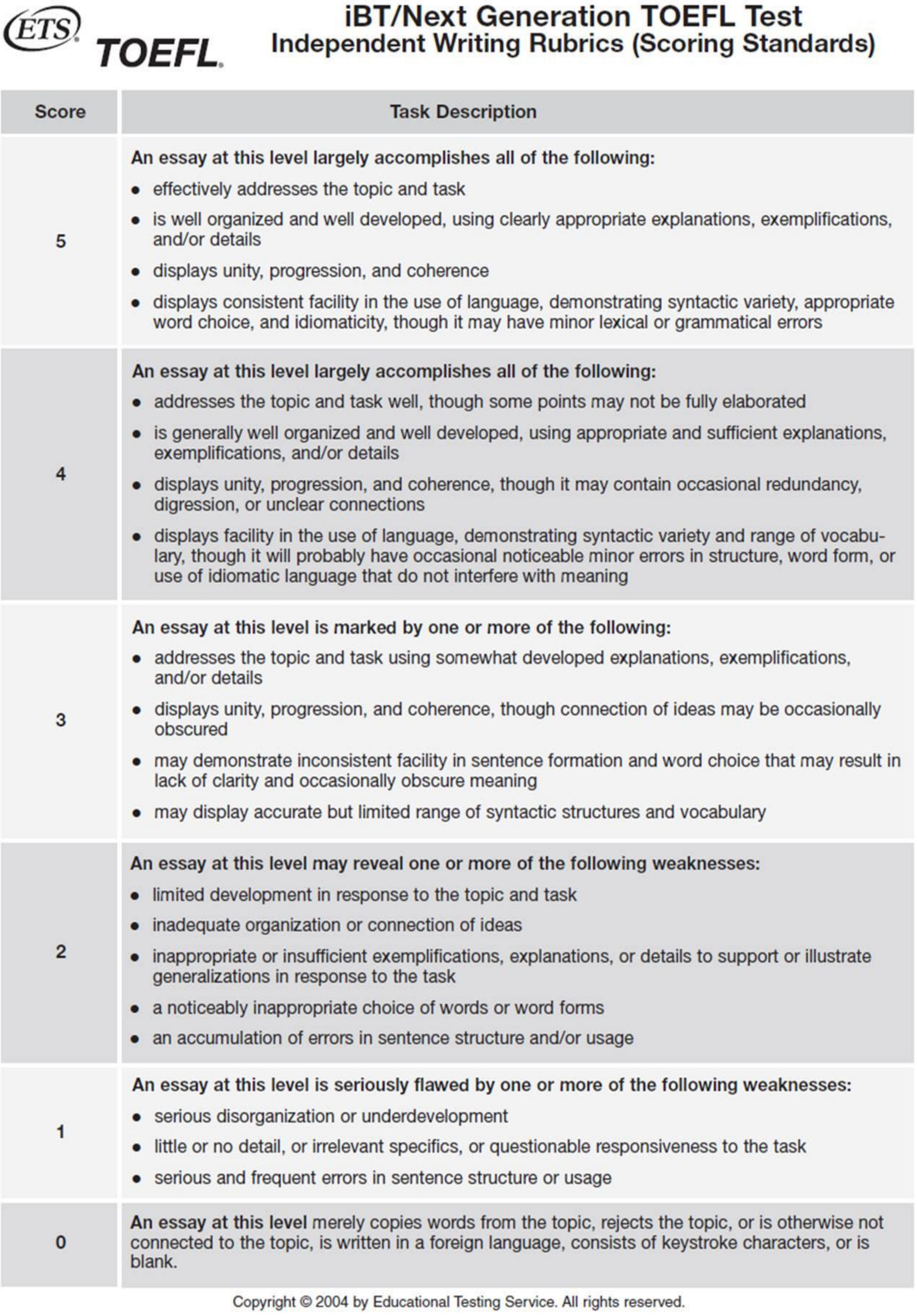


ETS TOEFL

iBT/Next Generation TOEFL Test
Integrated Writing Rubrics (Scoring Standards)

Score	Task Description
5	A response at this level successfully selects the important information from the lecture and coherently and accurately presents this information in relation to the relevant information presented in the reading. The response is well organized, and occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections.
4	A response at this level is generally good in selecting the important information from the lecture and in coherently and accurately presenting this information in relation to the relevant information in the reading, but it may have minor omission, inaccuracy, vagueness, or imprecision of some content from the lecture or in connection to points made in the reading. A response is also scored at this level if it has more frequent or noticeable minor language errors, as long as such usage and grammatical structures do not result in anything more than an occasional lapse of clarity or in the connection of ideas.
3	A response at this level contains some important information from the lecture and conveys some relevant connection to the reading, but it is marked by one or more of the following: • Although the overall response is definitely oriented to the task, it conveys only vague, global, unclear, or somewhat imprecise connection of the points made in the lecture to points made in the reading. • The response may omit one major key point made in the lecture. • Some key points made in the lecture or the reading, or connections between the two, may be incomplete, inaccurate, or imprecise. • Errors of usage and/or grammar may be more frequent or may result in noticeably vague expressions or obscured meanings in conveying ideas and connections.
2	A response at this level contains some relevant information from the lecture, but is marked by significant language difficulties or by significant omission or inaccuracy of important ideas from the lecture or in the connections between the lecture and the reading; a response at this level is marked by one or more of the following: • The response significantly misrepresents or completely omits the overall connection between the lecture and the reading. • The response significantly omits or significantly misrepresents important points made in the lecture. • The response contains language errors or expressions that largely obscure connections or meaning at key junctures, or that would likely obscure understanding of key ideas for a reader not already familiar with the reading and the lecture.
1	A response at this level is marked by one or more of the following: • The response provides little or no meaningful or relevant coherent content from the lecture. • The language level of the response is so low that it is difficult to derive meaning.
0	A response at this level merely copies sentences from the reading, rejects the topic or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

Appendix II: TOEFL iBT Independent writing rubrics



ETS TOEFL

iBT/Next Generation TOEFL Test
Independent Writing Rubrics (Scoring Standards)

Score	Task Description
5	**An essay at this level largely accomplishes all of the following:** • effectively addresses the topic and task • is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details • displays unity, progression, and coherence • displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors
4	**An essay at this level largely accomplishes all of the following:** • addresses the topic and task well, though some points may not be fully elaborated • is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details • displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections • displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning
3	**An essay at this level is marked by one or more of the following:** • addresses the topic and task using somewhat developed explanations, exemplifications, and/or details • displays unity, progression, and coherence, though connection of ideas may be occasionally obscured • may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning • may display accurate but limited range of syntactic structures and vocabulary
2	**An essay at this level may reveal one or more of the following weaknesses:** • limited development in response to the topic and task • inadequate organization or connection of ideas • inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task • a noticeably inappropriate choice of words or word forms • an accumulation of errors in sentence structure and/or usage
1	**An essay at this level is seriously flawed by one or more of the following weaknesses:** • serious disorganization or underdevelopment • little or no detail, or irrelevant specifics, or questionable responsiveness to the task • serious and frequent errors in sentence structure or usage
0	**An essay at this level** merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.