

## Duolingo English Test: An Alternative Online English Proficiency Test

Heng Lu

Chulalongkorn University, Bangkok, Thailand

Corresponding author: [luhaheng@gmail.com](mailto:luhaheng@gmail.com)

Article information	
<b>Abstract</b>	The test view is on the Duolingo English Test (DET), an alternative online English proficiency test with a machine-driven characteristic. The review covers essential information of the DET such as test purpose, usage, score-mapping with CEFR scale, price, and publisher. Meanwhile, the test usefulness is discussed with focuses on reliability, content validity, and impact. Both good points and weaknesses are illustrated. Furthermore, the view tries to criticize the AI algorithm issues in test content and score rating.
<b>Keywords</b>	Duolingo English Test, AI, algorithm, CAT, test review
<b>APA citation:</b>	Lu, H. (2023). Duolingo English test: An alternative online English proficiency test [Special issue]. <i>PASAA</i> , 66, 202–211.

## **1. Introduction**

The Duolingo English Test (DET) is an online English proficiency test, which has a distinct feature as a computer-adaptive test (CAT). It provides convenience for test takers in terms of test location choice with a flexible time. The test is created by language-learning platform Duolingo, who claims to have over 500 million users of its education app. In 2014, an eight-week research study was conducted by Ye (2014) focusing on the validity and reliability of the DET, however, Wagner and Kunnan (2015) pointed out that the research study, which was commissioned by Duolingo, lacked peer review. The test kept regular updates in test items. Meanwhile, the annually updated Duolingo English Test: Technical Manual, which was produced by Cardwell et al. (2023), contains useful information for both test takers and institutions about the test.

This test review begins with the test description, including information of test purpose, use, length, registration and retake, scoring-mapping, test edition, publisher, price, and score validity. Then, the CAT components are discussed, since this is the distinct feature of the DET. Regarding test usefulness, this review adopts qualities summarized by Bachman and Palmer (1996) with focuses on the validity, reliability, and some controversies of the AI-powered test.

## **2. Test Purpose**

The Duolingo English Test (2021) claims to measure English language competency for nonnative English learners in terms of communication and use in English language context. It assesses language users' ability in implementing language skills in four criteria, namely, literacy, conversation, comprehension, and production. Nevertheless, four skills of English language, namely listening, speaking, reading, and writing, can be indirectly represented via the test report in a combined score system.

### **3. Use**

The DET scores, which are interpreted as reflecting test takers' English language competency, can be used in post-secondary school admission applications as proof of English language proficiency. According to the official website of DET, the test scores are accepted by universities in over 50 countries, and the usage covers admission, placement test, progress test, and exit test. Meanwhile, the DET shares the same usages with another two international proficiency tests, namely the TOEFL iBT and the IELTS Academic.

### **4. Length**

The test length of the DET is one hour. The test taker will be guided through three different sections, which include i. Introduction and onboarding (5 minutes); ii. Adaptive test (45 minutes); and iii. Writing sample and speaking sample (10 minutes). However, the introduction and onboarding section cannot be regarded as test content since it simply aims to conduct test takers' hardware adjustment, identification document check, and an introduction of the test administration rules.

### **5. Registration and Retake**

Test takers have to register for the DET online at the official website at <https://englishtest.duolingo.com/register>, and relevant test information is available online. There is no limit for test takers to retake the DET, but the new test will only be available after the completion of the former test results, which is normally within 48 hours of the test submission.

### **6. Score: Mapping the DET Score to CEFR vs. School's Interpretation**

The point system used by DET changed in mid-2019. Prior to that, the DET reported a 100-point scale. The 160-point scale has been adopted since July 15, 2019, and the increment has also changed accordingly. A 1-point increment was used under the 100-point scale system, while a 5-point increment is currently in

use under the 160-point scale. The test explains that the 160-point scale, ranging between 10 to 160, can help distinguish itself from other tests.

The test mapped its scores to the Council of Europe Framework of References for Languages (CEFR) scale. The Duolingo English Test (2023) claims that its score range between 10 and 55 means the level of A1 to A2, 60 to 95 for B1, 100 to 125 for B2, 130 to 150 for C1, and 155 to 160 for C2. In contrast, the TOEFL iBT only mapped its scores from B1 to C2 level of CEFR, and the IELTS Academic mapped the scores from A2 to C2 of CEFR.

The point systems among these three tests are different. The Educational Testing Service [ETS] (n.d.) states that TOEFL iBT reports a score range from 0 to 120 with a 0.5-point increment from four subscores of listening, reading, writing, and speaking, while the IELTS Academic reports test score with the same four skills in a point system from 0 to 9 with an increment of 0.5 points. However, universities take a different way to interpret the DET score. For example, the Office of Admissions at the University of Washington (2023), in its test score requirement for English proficiency for international freshman applicants, states that a DET overall score of 105 points will be considered as the same level as IELTS Academic 6.0 points and TOEFL iBT 76 points. The university's interpretation of test score differs from the DET's description, which claims a strongly positive statistical relationship of score comparison between the DET - TOEFL iBT and the DET - IELTS Academic (Cardwell et al., 2023).

## **7. Home Edition vs. Test Center**

The DET is not available in test centers which means it is a home edition test. In contrast, both the TOEFL iBT and the IELTS Academic offer home edition plus test center versions.

## **8. Author and publisher**

The DET is developed and owned by a US-based Nasdaq-listed mobile learning platform Duolingo, Inc. The company offers more than 100 language courses in 41 languages on its website and app; however, the test has no explicit connection to the Duolingo language course content.

## **9. Price**

The price for taking DET is US\$49, far below the IELTS Academic of US\$245 (in the US) and the TOEFL iBT around US\$235 (test center in New York). The test taker can share the certified results to institutions or target schools unlimited times at no extra cost.

## **10. Score Validity**

The score report of the DET will be sent through the test takers' account. Like the IELTS Academic and TOEFL iBT, the DET result is valid for two years.

## **11. CAT Components**

The DET has 12 different item types for grading, which assess test taker's English language competency in an integrated manner in literacy, conversation, comprehension, and production (Cardwell et al., 2023). This is designed based on the theory of Cumming (2013), with an emphasis on the connection in integrated tasks between comprehension and production. Meanwhile, five item types of the DET belong to the CAT which means that the difficulty of questions will be influenced and changed depending on the previous question's answers. CAT related bias could occur since algorithm selection may overexpose some items in the test and completely ignore others (Chang & Van der Linden, 2003). However, CAT could use the algorithm to shorten the text content to provide an accurate pass or fail for test takers (Luecht, 2016).

## 12. Test Usefulness

In order to examine test usefulness of the DET, the review adopted the test usefulness components based on Bachman and Palmer (1996) with focuses on reliability, content validity, and impact.

### 12.1 Reliability

Regarding reliability, Ye (2014) explained the test-retest reliability coefficient for the DET was .79 (p. 7), pointing out the Duolingo test scores are reliable in the 100-point system. While in the new 160-point system, the test-retest reliability coefficient was .93, and those of the four subscores ranged between .90 and .92 (Cardwell et al., 2023).

The human examiner is not included in the DET throughout the test, which can be a positive aspect of test reliability. Meanwhile, when the AI test tries to assess speaking skills through the test items, the speaking scores are not presented in an independent separate sub-test. Instead, the DET evaluates speaking abilities through both conversation subscore and production subscore.

Another issue is that the speaking skills are assessed from five item types, namely (1) Listen and Select (CAT), (2) Read Aloud (CAT), (3) Speaking About the Photo, (4) Read, Then Speak, and (5) Listen, Then Speak, plus an ungraded speaking task called Speaking Sample. Could these item types be used to assess English speaking skills for academic purposes? This question needs to be answered. It is obvious that bias could be generated in the data preparation stage of the test, which is related to the data selection through the help of algorithm. Furthermore, although computer science can be used to design a system, it does ignore a lot of social context (Hao, 2019). Some bias examples, such as accent bias and local lingo bias, indicate that test score will be affected if the test takers use jokes or slang in the speaking test (IDP IELTS, n.d.).

Machine or human inaccuracy inevitably occurs in the score rating part. The IELTS Academic explains that a human examiner can understand the difference between a pause for thought and a pause for confusion from the test taker. It further emphasizes that only a human examiner, after receiving the specific training to put social context into consideration, can better serve the test. To clarify the issue, the DET explains that its test items undergo fairness and bias review by human raters to guarantee fairness for test takers of diverse identities and backgrounds (Cardwell et al., 2023).

In fact, the DET also utilizes human experts in its rating process. It describes that the tests, after being completed and uploaded, would undergo a thorough proctoring review by human proctors, which will be supplemented by AI. However, the DET human proctor differs from the examiner of its counterpart international proficiency test, namely the IELTS Academic, with regard to the speaking test. The human in the DET only takes part in assessing the examinee's test content, namely the audio and video files, while the latter test's human examiner is involved in both the real time face-to-face element of the speaking test and the scoring part. Since the speaking test requires a high level of authenticity between test takers and test content, the human-to-human test session can inherently provide a more authentic language speaking environment and knowledge content than the AI test.

## **12.2 Content validity**

Content validity of the test is an important issue. It indicates whether the test is composed of a representative sample of the language skills and structures that are designed to be measured (Hughes, 2002). The DET claims to measure test takers' English language proficiency in a general way instead of an academic way, however, the main usage of the test is as an English language proficiency proof for admission to universities and secondary schools. Wagner and Kunnan (2015) directly pointed out the gap of the language characteristics between academic English and general English. Thus, it seems inappropriate to regard the test as a

similar test to the TOEFL iBT and the IELTS Academic, which were purposefully designed to evaluate test takers' English language proficiency in academic context.

Another challenging part is whether the yes/no (text) tasks and the yes/no (audio) tasks of the DET can be used to assess the so-called ability in English reading and writing, or simply to assess the word vocabulary inventory of test takers. For example, whether the task of "select real English words from the list" could provide a clear assessing purpose towards test takers remains in question.

### **12.3 Impact**

Another issue is the impact concern. Although the computer-based test is accepted by many universities in the US, it did not expand much in Asia, and few Chinese public universities accepted the DET. While public universities in Thailand held a welcoming attitude towards the new test, for example, Chulalongkorn University accepted the DET score as one of the English proficiency tests for admissions into a number of its international programs such as the Bachelor of Arts in Language and Culture program, the Bachelor of Fine and Applied Arts program, and the Bachelor of Arts in Economics.

## **13. Conclusion**

It is evident that the AI featured test has many advantages in test time, scoring efficiency, location convenience, and low cost. However, the content validity issue and the lack of real human-to-human interaction still generates much concern about its test quality and validity. As the DET is continuously upgrading the test content and item types year by year, further improvement can be expected in the future.

Thanks to the wider recognition from schools worldwide, test takers can use the DET for college admission application as language proficiency proof. However, the schools using the test needs to consider how to interpret the score report in comparison with other international English language proficiency tests.

## 14. About the Author

Heng Lu is currently a PhD student of the English as an International Language Program at Chulalongkorn University in Bangkok, Thailand. He obtained a Bachelor of Arts in English Language, and a Master in Language and Culture from Linköping University. In 2023, he is also a part-time English lecturer at Chulalongkorn University Language Institute. His research interest is in language assessment and evaluation in EFL.

## 15. References

- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.
- Cardwell, R., Naismith, B., LaFlair, G.T., & Nydick, S. (2023) *Duolingo English test: Technical manual*. Duolingo English Test.  
[https://duolingo-papers.s3.amazonaws.com/other/technical\\_manual.pdf](https://duolingo-papers.s3.amazonaws.com/other/technical_manual.pdf)
- Chang, H. H., & Van der Linden, W. J. (2003). Optimal stratification of item pools in  $\alpha$ -stratified computerized adaptive testing. *Applied Psychological Measurement*, 27(4), 262–274.
- Cumming, A. (2013). Assessing integrated skills. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp.216–229). John Wiley & Sons.
- Duolingo English Test. (2021) Research and security: Demystifying online testing ‘mythbusting’ misconceptions and showcasing innovation [Webinar]. Duolingo. [https://www.youtube.com/watch?v=SyQoAQYn\\_4I](https://www.youtube.com/watch?v=SyQoAQYn_4I)
- Duolingo English Test. (2023). *Duolingo English test official guide for test takers*. [https://englishtest.duolingo.com/prepare/guide?landing\\_url=https%3A%2F%2Fenglishtest.duolingo.com%2Freadiness](https://englishtest.duolingo.com/prepare/guide?landing_url=https%3A%2F%2Fenglishtest.duolingo.com%2Freadiness)
- Educational Testing Service. (n.d.). *Setting TOEFL essentials score requirements*. <https://www.ets.org/toefl/score-users/essentials/set-score-requirements.html>
- Hao, K. (2019, February 4). This is how AI bias really happens—and why it's so hard to fix. MIT Technology Review.  
<https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/>
- Hughes, A. (2002). *Testing for language teachers*. Cambridge University Press.

- IDP IELTS. (n.d.). *IELTS Speaking test: AI vs face-to-face speaking*.  
<https://ielts.idp.com/prepare/article-ielts-speaking-ai-vs-face-to-face>
- Luecht, R. M. (2016). Computer-adaptive testing. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorisch, F. Ruggeri & J. L. Teugels (Eds.), *Wiley StatsRef: Statistics reference* online, (1–10). John Wiley & Sons.  
<https://doi.org/10.1002/9781118445112.stat06405.pub2>
- Office of Admissions at the University of Washington. (2023). *English proficiency for international freshman applicants*.  
<https://admit.washington.edu/apply/freshman/how-to-apply/english-proficiency/>
- Wagner, E., & Kunnan, A. J. (2015). The Duolingo English test. *Language Assessment Quarterly*, 12(3), 320–331.  
<https://doi.org/10.1080/15434303.2015.1061530>
- Ye, F. (2014). *Validity, reliability, and concordance of the Duolingo English Test*. School of Education, University of Pittsburgh.  
<https://docplayer.net/2636113-Validity-reliability-and-concordance-of-the-duolingo-english-test.html>