The Duolingo English Test — Design, Validity, and Value

duolingo english test

DET Whitepaper (Short)
September 28, 2017 (3 pages)
englishtest.duolingo.com/resources

Jeffrey Brenzel and Burr Settles

Introduction

Colleges, universities, and secondary schools around the world are using the Duolingo English Test (DET) as a new way to assess English language proficiency. The DET offers many advantages to admissions offices and applicants: on-demand accessibility, low cost, remote test proctoring, and rapid score reporting, as well as an integrated video interview and writing sample.

The DET also looks and functions differently from other proficiency tests, which raises an obvious question. If DET item types and administration differ from those used on other large-scale tests like the TOEFL® or IELTS®, can DET test scores be used in the same way as the scores from those instruments?[*]

New Objective, New Technology, New Design

Many innovations we now take for granted had to confront the initial challenge of unfamiliarity. When Amazon began selling books over the Internet, few people could imagine shopping online. Facebook created a tool for managing and sustaining a social network before most people recognized they had one.

When Duolingo launched its free language-learning app in 2012, we engineered an app to utilize short bursts of user time on mobile devices. Though no one had ever considered providing language instruction in this form, we quickly became the largest language-learning platform in the world, with more than 200 million learners in over 200 countries.

In 2014, we set out to design a new type of English proficiency test: one with greater efficiency, better security, lower cost, and universal access. To succeed, we knew the test would require a new and unfamiliar design.

The Duolingo English Test Construct and Design

The DET testing protocol combines practical online delivery and remote proctoring with the following advances:

- A statistical machine learning approach to creating *criterion-referenced* assessment constructs based on the Common European Framework of Reference (CEFR),
- Innovative, research-driven test item formats that can be *generated in large quantities*, and
- An efficient *computer-adaptive testing* (CAT) protocol.

The overall test construct of the DET is calibrated to the Common European Frame of Reference, or CEFR (Council of Europe, 2001). This is a global standard used to characterize multiple levels of language proficiency. Specifically, we use machine learning and natural language processing technology to fit statistical models to tens of thousands of CEFR-annotated text passages and wordlists. The overall test development process is illustrated in Figure 1.

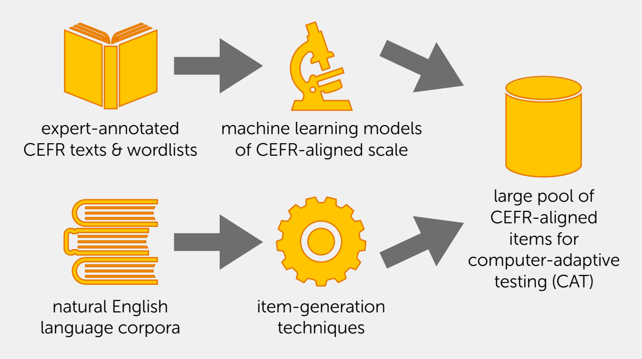

expert-annotated CEFR texts & wordlists
machine learning models of CEFR-aligned scale
large pool of CEFR-aligned items for computer-adaptive testing (CAT)
natural English language corpora
item-generation techniques

Figure 1. Illustration of the DET construct and test item creation.

Correlations between these machine-learned CEFR models with human judgments are very high (between 0.65–0.95, depending on the test item type), which is good evidence for construct validity. In this way, the DET is capable of generating and calibrating a large number of test items, while still achieving inter-item and test-retest reliability metrics that exceeds industry standards (Settles, 2016). The large item pool further enhances test security by virtually eliminating the chance that any two test administrations will share the same items.

Furthermore, the DET does not rely upon multiple-choice item formats that provide limited information about the examinee's ability, but rather focuses on interactive items (such as listening transcription and speaking exercises) that can be generated in large quantities and automatically scored. We have mined the language assessment research literature to select item types that combine simple and intuitive formats with multiple compact measurements and strong predictive capabilities.

For example, one item type requires examinees to discriminate between real English words (e.g., *meeting* and *thunder*) and English-like pseudowords (e.g., *clerm* and *earts*). Though this is not obviously related to everyday language tasks, it does require the same cognitive processes (e.g., lexical and morphological activation) that are used in everyday reading, writing, and even listening activities. Hence, it has been demonstrated for decades that this unusual but engaging assessment format significantly predicts all three of these language skills (Milton, 2010; Staehr, 2008; Zimmerman et al., 1977).

[*] This "living document" is periodically updated, as the DET platform advances and new research becomes available. This version is dated September 28, 2017. See `https://englishtest.duolingo.com/resources` for the latest version.

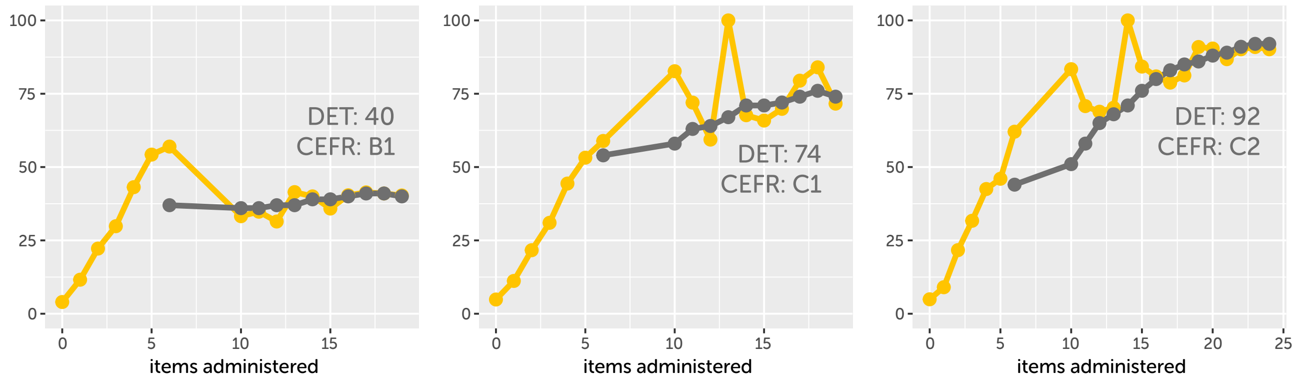


Figure 2. Example computer-adaptive testing (CAT) sessions from the DET. As test sessions progress, the CEFR-aligned difficulty for each new item (yellow) is selected based on the examinee's estimated proficiency at that point in the test (gray). These estimates are updated after each item to adaptively select the next one (following an initial calibration set of 5–6 test items). This process can efficiently determine the examinee's overall score, and improve test security by making each administration virtually unique.

Similarly, Duolingo's scoring technology for speaking relies not only on basic computer speech recognition, but also on the acoustic properties of speech like intonation, rhythm, and stress. Such "suprasegmental" features have been shown to be primary contributors to conversational intelligibility (Hahn, 2004; Anderson-Hsieh et al., 1992) and are linked in turn to overall speaking ability (Derwing and Munro, 1997).

Importantly, the DET is a computer-adaptive test (CAT), meaning that the difficulty of test items presented within each session fluctuates based on how the examinee has performed on previous items. If examinees are doing poorly, they will see more basic test items; if they are doing well, they will see more challenging ones. See Figure 2 for examples of the adaptive protocol in action for three actual test cases of increasing ability.

CAT technology is commonly used in several other high-stakes tests, such as the GRE and GMAT. It not only significantly shortens test-taking time (Weiss and Kingsbury, 1984), but also provides greater test security, because each testing session presents the examinee with a unique path through a very large database of test items.

Test Validity

Unlike the most popular English proficiency tests, the DET relies less on tasks that appear to mimic student activities in a typical college course. For example, the DET does not incorporate a scored writing section at this time—though it does contain items known to predict writing ability, and also includes a writing sample for direct inspection by admissions offices. Similarly, the test does not ask examinees to answer questions about long reading passages—though several item types are known to assess discourse-level reading comprehension.

In other words, the test does not incorporate some tasks that appear to parallel tasks that one might encounter in a typical college course. However, DET items are strongly predictive of performance on these classroom tasks. Attempting to mimic those tasks more directly does not in itself ensure that a test will better predict holistic language functions.

The research conducted to date has produced evidence that the DET scores are substantially correlated with other popular English test scores. As shown in Figure 3, correlations between the DET and both the TOEFL and IELTS are above 0.70 and statistically significant ($p < 0.0001$). Meanwhile, the most recent study relating these two tests' scores to each other reports a correlation of 0.73 (ETS, 2010). This indicates that the DET test construct is aligned with TOEFL and IELTS about as well as they are aligned with each other.

Perhaps more importantly, recent research has indicated that the DET may even outperform more popular tests where it counts the most: predicting how the ESL faculty who support international students on the college campus will rate the English proficiency and the support needs of incoming international students (Ishikawa et al., 2016).

Duolingo is committed to continuing its own research and collaborating with partner schools and independent researchers to further validate the DET and substantiate previous findings.

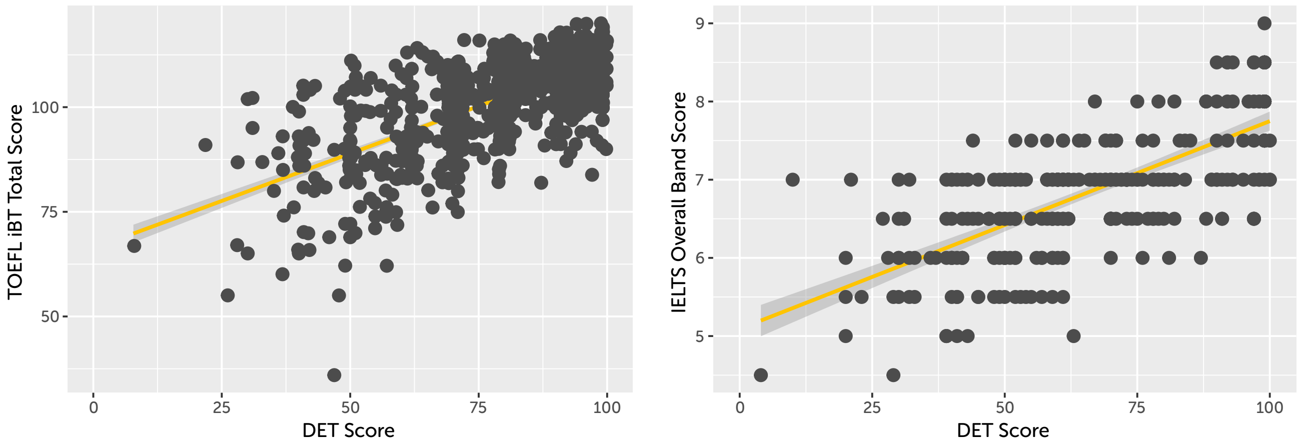
Integrated Interview and Writing Sample

In another marked departure from traditional testing protocol, Duolingo has added an integrated video interview and writing sample to the testing platform. We noted that given diminished confidence in traditional tests, many colleges and universities have opted to supplement standardized test scores by requiring some or all candidates to sit for video recorded interviews. These interviews provide a check on examinee identity as well as an opportunity for admissions officers to view candidates producing English in live interactions.

Unfortunately for candidates, the additional cost required for interviewing services (such as InitialView or Vericant) can be as high or higher than the cost for the standardized testing whose weaknesses the interviews are meant to address.

Because each DET session is fully video recorded, we are able to include a supplementary video recording for admissions offices at no additional cost to applicants. At the end of the scored test session, the DET presents four short, open-ended prompts (oral, written, and image formats) to the candidate, requiring that they produce spontaneous, 90-second responses. The examinee's performance on the scored portion of the test determines the nature and difficulty of the prompts with respect to their CEFR reference level. The DET score report includes not only the standardized test score, but also a direct check on speaking abilities, suitable for rapid review by time-pressured admissions officers.

In the newest release of the DET for the 2017–2018 admissions cycle, we have also integrated the collection of short writing



(a) TOEFL iBT ($r = 0.71$, $n = 1{,}057$) **(b)** IELTS ($r = 0.70$, $n = 304$)

Figure 3. Correlation of DET scores with (a) TOEFL iBT and (b) IELTS. These results are based on anonymized student data provided by partner institutions, where scores for both test were available for high-stakes admissions or ESL student placement.

samples. The writing sample consists of several open-ended image description tasks designed to elicit a range of vocabulary, grammatical, and syntactic abilities, as well as a longer-form essay in which examinees state an opinion, belief, or preference and defend it. Similar to the video interview, the writing prompt difficulties are based on performance on the other portions of the adaptive test. While writing samples are currently unscored, they are provided alongside the score report as a further check on written English skills.

Commitment to Continuous Improvement

Duolingo is a leader in educational technology, and we are committed to providing a language proficiency tool that meets the changing needs of examinees, admissions offices, and institutions. In the short time since its initial launch, the Duolingo English Test has already been updated and improved based on ongoing research and stakeholder feedback, and it will continue to evolve with new technology for each admissions cycle.

Creating a new kind of English proficiency test is an audacious task. It requires not only an effective and reliable solution, but also a paradigm shift in how we assess language ability, and a change in the established workflows of admissions offices. The Duolingo English Test may feel unfamiliar at first glance, but every aspect of the tool's construction has faced rigorous scrutiny to ensure that it is accurate, reliable, useful, and consistent with Duolingo's mission of making education accessible to all.

Author Biographies

Jeffrey Brenzel is the University Admissions Liaison for Duolingo. He was an administrator and teacher at Yale University for more than twenty years, including eight years at the Dean of Undergraduate Admissions from 2005–2013. He has engaged extensively with issues surrounding the design and use of testing for college admissions.

Burr Settles is one of the principal developers of the Duolingo English Test. He holds a PhD in computational linguistics and machine learning, and was a postdoctoral research associate at Carnegie Mellon University from 2009–2013 until he joined Duolingo. He currently leads the Duolingo research group applying large-scale computational methods to language learning and assessment research.

References

J. Anderson-Hsieh, R. Johnson, and K. Koehler. The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentais, prosody, and syllable structure. *Language Learning*, 42(4):529–555, 1992.

Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

T.M. Derwing and M.J. Munro. Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19(1):1–16, 1997.

ETS. Linking TOEFL iBT scores to IELTS scores—a research report. 2010.

L.D. Hahn. Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38:201–223, 2004.

L. Ishikawa, K. Hall, and B. Settles. The Duolingo English Test and academic English. Technical Report DRR-16-01, Duolingo, 2016. http://bit.ly/29WsDcu.

J. Milton. The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin, and I. Vedder, editors, *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, pages 211–232. Eurosla, 2010.

B. Settles. The reliability of Duolingo English Test scores. Technical Report DRR-16-02, Duolingo, 2016. http://bit.ly/2bpqLx9.

L.S. Staehr. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36:139–152, 2008.

D.J. Weiss and G.G. Kingsbury. Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21:361–375, 1984.

J. Zimmerman, P.K. Broder, J.J. Shaughnessy, and B.J. Underwood. A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence*, 1(1):5–31, 1977.