Duolingo English Test, Revised Version July 2019

Elvis Wagner

To cite this article: Elvis Wagner (2020): Duolingo English Test, Revised Version July 2019, Language Assessment Quarterly, DOI: 10.1080/15434303.2020.1771343

To link to this article: https://doi.org/10.1080/15434303.2020.1771343

Published online: 28 Jun 2020.
Duolingo English Test, Revised Version July 2019

Elvis Wagner

Teaching and Learning Department, Temple University, Philadelphia

ABSTRACT

The Duolingo English Test (DET) is a computer adaptive test of English proficiency that is increasingly used for English-medium university admissions purposes. During the COVID-19 pandemic of 2020, test centers were shut down in many countries, and major tests including the TOEFL iBT and IELTS could not be administered. The DET is an “at home” test, and thus many universities began accepting DET scores as a measure of applicants’ English proficiency. Because a revised version of the DET was launched in July, 2019, and because of the large increase in universities accepting DET scores, a critical review of the DET is warranted. The current review lauds the accessibility of the test (e.g., it is an inexpensive “at home” test that can be taken anywhere, in less than an hour, with scores returned in 48 hours). However, the test has multiple shortcomings: the test tasks have little in common with the types of language tasks university students engage in; the test does not assess test takers’ academic language ability, discourse level competence, or interactional competence; it is susceptible to cheating and test preparation; and it has a potential for negative washback on learners and learning systems. In addition, there is a lack of independent research validating the use of DET scores for admissions. Given these shortcomings, the use of DET scores cannot be recommended for university admissions purposes.

Introduction

In early 2020, as the COVID-19 crisis intensified, governments imposed restrictions to prevent the spread of the disease, including the closing of test centers in many countries. Prominent English proficiency tests that are used for university admissions purposes (including the TOEFL iBT and IELTS) were affected, resulting in many thousands of students hoping to study at an English-medium university not being able to provide English test scores as part of their university applications. To address this void, many universities began to accept the Duolingo English Test (DET) for admissions purposes, because the DET does not need to be administered at a test center.

While the DET officially launched in 2014, it was not widely accepted until the COVID-19 crisis of 2020. In 2015, I was the first author of a review of the DET (Wagner & Kunnan, 2015) that was published in Language Assessment Quarterly. Because of the recent increase of its acceptance by universities for admissions purposes, and because a modified version of the DET was introduced in 2019, I decided to write another review of the test for this journal.

The test

The Duolingo English Test is a computer-adaptive test (CAT) of English language proficiency that is administered and scored totally online. According to LaFlair and Settles (2019), scores on the DET reflect the English language ability of test-takers, and these scores are used for many purposes, “including for university admissions decisions” (p. 3). For this review, I will focus on the use of DET scores for admissions decisions for English-medium universities.

Administration and scoring

Because the test is administered totally online, test-takers do not need to travel to a test center; instead, they are able to take the test on any internet-connected computer, at any time (test-takers are advised that they need a government ID, a quiet, well-lit room, 60 minutes, a reliable internet connection, and a computer with a supported browser, front-facing camera, microphone, and speakers) (https://englishtest.duolingo.com/applicants). The DET is a computer-adaptive test (CAT), meaning that the difficulty level of the test tasks (or testlets) that test-takers receive varies based on the test-takers’ performance on the earlier part of the test. Higher scores on previous tasks result in more difficult ensuing tasks, while lower scores on earlier tasks result in easier ensuing tasks. In addition, the type of test task that test-takers receive is determined by their performance on the previous part of the test, and during a single test, a test-taker might repeat a particular test task type multiple times. CAT allows for a more efficient assessment, as fewer items are needed to assess test-takers’ ability. As a result, the DET is much shorter in duration (less than 45 minutes) than other standardized English proficiency tests (e.g., the TOEFL iBT and the IELTS take about three hours).

All of the scoring is done by computer, and test-takers get their certified results within 48 hours. Test-takers receive a single holistic score on a scale of 10–160. A test-taker who scores from 10 to 55 is considered to be at the CEFR levels A1-A2; a score from 60 to 115 aligns with CEFR levels B1-B2; and a score from 120 to 160 aligns with CEFR levels C1-C2 (https://englishtest.duolingo.com/scores).

Test tasks

There are seven different scored test tasks on the DET: C-test, Yes-No Vocabulary Task – Written, Yes-No Vocabulary Task – Oral, Dictation, Elicited Imitation (Read-Aloud), Extended Speaking, and Extended Writing. In addition, there are two unscored test tasks that are given at the end of the test: Oral Interview and Writing Sample. The first five tasks are CAT tasks, in that the difficulty levels of the tasks that are presented to test-takers depend on how they did on the previous part of the test. The difficulty levels of the items in these five test tasks are not based on previous test administrations. Rather, they are estimated through machine learning and natural language processing, based on a number of psycholinguistic variables, including syntactic, morphological, and lexical variables (see LaFlair & Settles, 2019; Settles, LaFlair, & Hagiwara, 2020 for more information). In addition, the task type presented will vary based on test performance up to that point. The technical manual for the test states that for each test administration, an individual test-taker will receive at least three of each of these task types, with a maximum of seven instances of an individual task type (with the median of six occurrences of each task type for the CAT portion of the test). This suggests that the average test-taker will be presented with 20–30 of these five CAT test tasks.

The Extended Speaking and Extended Writing tasks are not computer adaptive but do vary in difficulty, again based on the test-taker’s performance on the test up to that point.
Finally, the two unscored test tasks (Oral Interview and Writing Sample) are presented at the end of the test. According to the DET website, these tasks are not scored but are provided to the institutions that receive the test-taker’s results (https://testcenter.zendesk.com/hc/en-us/articles/360030587212-How-does-the-video-interview-and-writing-sample-work-).

These test tasks are described below, along with a sample demonstrating each task format.

a. C-test

The C-test task has a short written passage (three or four sentences), with the first and last sentence complete. The middle sentences have parts of words missing, with a blank for each missing letter of the word, and the test-taker must fill in the missing letters. LaFlair and Settles (2019) describe the task as having words that “… are ‘damaged’ by deleting the second half of the word. Test-takers respond to the c-test items by completing the damaged words in the paragraph” (p. 6). In the sample shown below, the “damaged” words are in text boxes, and the provided letters in the damaged words are in gray, and the letters that have been completed by the test-taker are shown in orange.

![C-test sample](image)

(LaFlair & Settles, 2019, p. 7)

b. Yes/No Vocabulary Task – Written

With this task, the test-taker is presented with a list of English words and pseudo-words (see LaFlair & Settles, 2019, for an explanation of how the pseudo-words are created). The test-taker must identify the actual English words by clicking on them. The total number of words and pseudo-words varies for each task.
c. Yes/No Vocabulary Task—Audio

In the audio version of the Yes/No Vocabulary task, the test-taker clicks on a sound icon, and an oral recording of the word is played. The test-taker then clicks on a check mark icon to highlight it if it is an actual English word and does not click/highlight it if it is not an actual English word. Again, the total number of words and pseudo-words that the test-taker must review varies for each task.
d. Dictation

In the dictation task, the test-taker hears a spoken text (the length varies according to the difficulty level of the task, but usually about a sentence or two), and then must type it using a keyboard. The test-taker has one minute to complete the task and can hear the spoken text up to three times.

(LaFlair & Settles, 2019, p. 9)

e. Elicited Imitation (Read-aloud)

In the elicited imitation task, the test-taker is presented with a written sentence and then must speak the sentence aloud while using the computer microphone.

(LaFlair & Settles, 2019, p. 9)
f. Extended Speaking

The extended speaking task requires a test-taker to respond orally to four prompts: a picture description, two different written prompts, and an aural prompt (test-takers can hear the aural

(LaFlair & Settles, 2019, p. 11)
prompt a maximum of three times). The difficulty level of these tasks (high, intermediate, or low proficiency) is determined by the test-taker’s performance on the previous portion of the test. According to the sample test provided at https://englishtest.duolingo.com/applicants, the test-taker is asked to speak “for at least 30 seconds” in response to the aural prompt. For at least some of the prompts, the test-taker is given 30 seconds to prepare a response to the prompt.

a. Extended Writing

For the extended writing task, the test-taker responds to four different prompts: three picture description prompts and a written prompt in which the test-taker is required “to describe, recount, or make an argument” (LaFlair & Settles, 2019, p. 10), and the response must be at least 50 words. As with the extended speaking tasks, the writing prompts are of varying difficulty levels, and the difficulty of the prompt that a test-taker receives is based on his/her performance on the computer-adaptive portion of the test.

(LaFlair & Settles, 2019, p. 12)
g. The Oral Interview and Writing Sample

At the end of the DET, the test-taker completes an oral “interview” and also completes a writing sample. The prompts for these tasks advise the test-taker that, “Your responses will not affect your score but will be available to institutions that receive your score”. With the oral interview, the test-taker is given a choice of two topics to reply to. The test-taker gets 30 seconds to choose a question to respond to, and then must speak for one to three minutes on the topic, while being videorecorded. With the writing sample, the test-taker is also given a choice of two topics to reply to (and 30 seconds to choose), and then three to five minutes to write about the topic. Again, these responses are not scored but are shared with the institutions that receive the test-takers' scores from Duolingo.

Test qualities

In this section, I provide my analysis and critical review of the DET, focusing on the test qualities of accessibility, reliability, validity, potential for cheating, and impact, in relation to its use for English-medium university admissions purposes.

Accessibility

The accessibility of the DET is its obvious strength. The test was designed specifically to be accessible, efficient, and user-friendly (LaFlair & Settles, 2019). It does not require a trip to a test center, but can be taken anywhere with a computer with an internet connection. It is much less expensive than other prominent English proficiency tests (at 50 USD, only one-fourth or one-fifth the cost of competitors), and the entire onboarding and testing process can be completed in less than one hour. Test-takers receive their score results in less than 48 hours. The value of this increased accessibility became very clear during the COVID-19 crisis of 2020 when tests such as TOEFL iBT and IELTS were forced to close testing centers for an extended period. Indeed, the TOEFL iBT and IELTS soon developed “at home” versions of their tests.

Another component of accessibility is related to accommodations made for test-takers with disabilities. On the DET webpage (https://testcenter.zendesk.com/hc/en-us/articles/360035992711-Can-I-request-an-accommodation-for-a-disability-), test-takers can request an accommodation for a disability by emailing with an accommodation request, but this seems to be the only information provided, and there does not seem to be any mention of this in the DET research literature, nor any information about what sorts of accommodations are provided.

Reliability

Reliability of scoring would also seem to be a strength of the DET, in that it is entirely computer scored, and thus scoring error due to human characteristics (e.g., fatigue, bias, differing interpretation of rubrics, etc.) is avoided. For the extended speaking and writing tasks, LaFlair and Settles (2019) describe how the algorithms used for scoring these tasks include grammatical accuracy, grammatical complexity, lexical sophistication, lexical diversity, task relevance, and length. The scoring algorithm for the speaking tasks also includes fluency and acoustic features. They also state that the algorithms were evaluated using “cross-validation” and that human-machine agreement (using Cohen’s κ as the agreement index) was higher (.73 for writing, .77 for speaking) than human-human agreement (.68 for writing, .77 for speaking) (LaFlair & Settles, p 15).

One concern about the reliability of the test is due to the huge number of items that are required for a CAT test. It is unclear if the DET includes some sort of bias review for the items. Testing organizations usually have some sort of bias review to make sure that test-takers do not receive tasks or items that could be inappropriate or offensive.
Another concern regarding reliability involves the algorithm used in scoring the speaking tasks. With a human rater, bias is possible during the scoring process. A rater who has a (conscious or unconscious) bias towards a particular accent variety might score the speaking lower because of that bias, introducing error into the scoring. However, while machine scoring would seem to reduce this particular error in rating, there is still the question of what aspects of the speech (the “acoustic features”) are actually being measured, and the possibility that a particular accent might be scored more harshly than a different accent variety, even if they were equally intelligible.

**Validity**

The focus of this review is the use of DET scores for English-medium university admission purposes. When reviewing validity, then, the central focus is whether DET scores allow university admissions officers to make valid inferences about whether test-takers have the necessary level of English ability that will allow them to be able to perform in an academic university context. Validity, of course, is not a yes/no proposition; instead, a test developer has to make a coherent and convincing argument (based on empirical data) that its test is valid for the particular purpose it is being used for. In reviewing the DET, there are a number of obvious areas in which the test is problematic, indicating that the appropriateness of using this test for university admissions purposes is suspect.

**No assessment of discourse-level competence**

A shortcoming of the test is that it does not seem to be assessing test-takers’ ability to create or comprehend discourse-level English. Obviously, in a university setting, students need to use language to comprehend and produce texts of varying lengths: they read books and articles, write papers, give oral presentations, listen to academic lectures, etc. Yet the DET does not assess test-takers’ ability to produce or comprehend texts beyond the sentence (or at most, paragraph) level. Of the seven scored tasks used on the test, two of the tasks (Yes-No Vocabulary Task – Written, Yes-No Vocabulary Task – Oral) are at the word level; two (Dictation, Elicited Imitation) are at the sentence level; and three are at the paragraph level (C-test, Extended Speaking, and Extended Writing). Even the “extended” speaking and writing tasks are actually quite short, with test-takers told to respond with at least 50 words in writing, or for at least 30 seconds in speaking. Only the oral interview and writing sample tasks involve extended discourse, although even these tasks are relatively short (the oral interview is one to three minutes, and the writing sample asks the test-taker to write for three to five minutes). But again, the oral interview and writing sample are not actually scored.

**No assessment of interactional competence**

The idea of interactional competence has been a fundamental tenet of second language acquisition for decades (Kramsch, 1986), and interactional competence is one of the underlying constructs of L2 assessment. The goal of language assessment is widely seen as assessing a test-taker’s ability to interact successfully with others in the target language, both orally and literately, and appropriately within the context of the situation. This is a broad construct, and obviously test developers are not able to assess all aspects of this construct. Instead, they are forced to choose language tasks that are representative of the target language use (TLU) domain of interest. The TLU domain of interest here is an English-medium university setting, in which students are expected to be able to interact orally with fellow students, professors, staff, etc. They are expected to read written texts (e.g., books, articles, essays) and listen to spoken texts (e.g., lectures, arguments, explanations), and then do something with that information (e.g., engage in an interactive conversation, participate in a class discussion, write an academic paper based on learned information, create an oral report, take a test, etc.). It is this idea of interaction, with the language, with language texts, and with other speakers, in a sociolinguistically appropriate manner for that particular context, that undergirds the notion of interactional competence as a fundamental component of language ability.
But the DET seems to lack any real attempt to measure test-takers’ interactional competence. The two Yes-No Vocabulary tasks involve no interaction, nor do the dictation and elicited imitation tasks. These tasks require no actual comprehension of language, and certainly no interaction based on language comprehension. The extended speaking and extended writing tasks require test-takers to respond to a picture or a single question. The C-test involves minimal interaction.

One of the common criticisms of large-scale standardized tests of English proficiency is that they rely (too much) on the use of multiple-choice comprehension questions. Brenzel and Settles (2017) state that, “the DET does not rely upon multiple-choice item formats that provide limited information about the examinee’s ability” (p. 1). This is a fair critique, in that there is real concern about the ability of these types of questions to actually tap into a test-taker’s interactional competence. But Brenzel and Settles go on to state that the DET instead uses “interactive items (such as listening transcription and speaking exercises) that can be generated in large quantities and automatically scored” (2017, p. 1). The argument seems misguided at best, if not disingenuous. The test tasks on the DET are not interactive and do not assess test-takers’ interactional competence.

**Lack of authentic test tasks**

Bachman and Palmer (1996) define authenticity in the context of language testing as the “degree of correspondence of the characteristics of a given language test task to the features of a TLU (target language use) task” (p. 23). The test tasks on the DET have very little correspondence with the types of language tasks that university students would typically engage in. Authenticity is not a requirement for a language test, but having authentic test tasks is useful for a number of reasons: using authentic test tasks should allow for more valid inferences about test-takers’ ability beyond the test setting; it can have a positive washback effect; it should make cheating more difficult; and test-taker affect and motivation should be improved when the test-taker can see that the assessment tasks are similar to real-world language tasks (Ockey & Wagner, 2018).

In examining the DET test tasks, it is obvious that the tasks are dissimilar to and not representative of the language tasks found in a university TLU domain. Indeed, Brenzel and Settles (2017) describe how the DET “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” (p. 2). (This article was published two years before the test was revised into its current state, but the revised test tasks still do not “appear to parallel tasks” that are found in most university courses.)

An example of the inauthentic nature of the test tasks is the use of the two Yes/No vocabulary tasks (written and oral) that requires test-takers to identify actual English words, as opposed to pseudo-words. These tasks have virtually no correspondence with the types of language tasks common in a university setting. Brenzel and Settles (2017) admit that while this type of task “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, Broder, Shaughnessy, & Underwood, 1977)” (p. 1). But is this really true? These Yes/No vocabulary tasks are almost never used in L2 assessment, and rarely in SLA research. A closer examination of the studies cited here do little to support the use of these tasks for assessing L2 proficiency. The Zimmerman et al. (1977) study is a psychology study that used Yes/No vocabulary tasks to examine the correlation of L1 vocabulary knowledge with SAT scores, but it had nothing to do with L2 learners or assessment. The Staehr (2008) study did find that L2 vocabulary knowledge correlates with reading, writing, and listening (certainly not surprising results), but the Staehr study did not use Yes/No vocabulary tasks. Milton (2010) is a review of studies examining vocabulary breadth. Milton does argue for the use of Yes/No vocabulary tasks to measure the breadth of L2 vocabulary knowledge, but there is no validation of their use for this
purpose. Brenzel and Settles (2017) claim that the DET developers “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” (p. 1), but the test tasks used on the DET are rarely found in the language assessment research literature. Instead, it seems more likely that the need to generate and automatically score large quantities of items is dictating the test design, rather than an attempt to truly assess test-takers’ interactional competence in English.

What exactly are these Yes/No vocabulary tasks measuring? In taking the DET as part of the preparation for this review, I had to perform both the written and oral versions of this task many times over, seemingly more than all the other tasks combined. How much of the overall score of the test is contributed by these tasks? In addition, it would seem that these tasks could be susceptible to cheating. It would have been possible for someone else to have been in the room while taking the test, and I could read out the words and pseudowords, and the other person (in my line of sight, but not visible in the webcam) could signal “yes” or “no”. In addition, this task would also seem to be particularly susceptible to test preparation. Future test-takers could study lists with hundreds or thousands of English words, memorizing the orthographical form of the word, without knowing the meaning or anything else about it, and still score well on these tasks.

But this lack of authenticity goes beyond the Yes/No vocabulary tasks. At first glance, dictation might seem to be a type of language task common in a university setting, but this dictation task does not require the listeners to actually comprehend the spoken text; rather, test-takers just have to be able to type it out. And in most real-life listening situations, listeners do not listen to a text word for word, as is required with the dictation task. Similarly, the elicited imitation task does not require comprehension, but only the ability to pronounce accurately while reading aloud. The C-test tasks are not language tasks university students are expected to do. Only the extended speaking and extended writing tasks have characteristics somewhat similar to the language tasks university students would typically perform, but even these tasks are hardly similar to the types of activities students would do in class, including comprehending (extensive) spoken or written information, analyzing it, and transforming it in some way. These test tasks involve almost no comprehension of spoken or written texts, no assessment of discourse competence, and no assessment of interactional competence.

On the DET website, under the heading “Score Interpretation”, it is stated that “The Duolingo English Test is designed to assess real-world language ability, by measuring reading, writing, speaking, and listening scores” (https://englishtest.duolingo.com/scores), but this statement seems particularly disingenuous, as the language ability that is being assessed on the DET is almost nothing like the real-world language ability required of university students.

**Potential for cheating**

There have been a number of cheating scandals related to large-scale English proficiency tests, and test organizations seem particularly aware of and sensitive to the issue. The DET touts their procedures for preventing cheating, including requiring test-takers to enable the webcam on their computer so that the session can be recorded, requiring test-takers to have a valid government ID that is photographed as part of the onboarding process, and a list of rules for the test-taker (must be alone in a well-lit room, ears visible and uncovered, face visible and unobscured, no head/earphones, no outside materials, no communicating with anyone else, no writing utensils or paper, no looking away from the screen). While taking the test there is a “browser lockdown” that results in the test being cancelled if test-takers try to use other applications or plugins. The DET assures institutions that they have “rigorous test proctoring” that involves “Expert human proctors, with the help of AI, examine each test session for more than 75 different behaviors in multiple, independent rounds of review” (https://englishtest.duolingo.com/security). Finally, the videorecorded interview and writing sample are shared with the institutions that receive the scores.
Many of these measures would seem to be effective, in both catching cheating and also acting as a deterrent to prevent it. However, in high-stakes testing, it is a truism that the higher the stakes, the more innovative the cheater. The fact that the test-taker can take the test at home, rather than at a test center, would seem to make the chance of getting caught and possibly punished much less successful as a deterrent. One of the most obvious ways to cheat on any high-stakes test is to have someone take the test for you, using your own government identification. If someone looks vaguely like you, that person could take your ID, hold it up to the camera, and then take the test. Obviously, this can happen in a test center as well, but the stakes there would seem to be much higher (i.e., giving your ID to someone else, the imposter having to physically go to and hand over or show the borrowed ID, etc.). If an imposter is caught at a test center, the ID could be impounded, or even more severely, the person could be detained. Using a borrowed or falsified ID on an “at home” test, however, would seem to have much less severe consequences, and thus the risk of being caught would seem to have much less of a deterrent effect.

Test developers understandably are reluctant to acknowledge cheating and allow independent researchers to examine its prevalence. But because “at home” tests seem to be more susceptible to cheating, the issue needs to be addressed through research. This issue is not unique to DET, of course, now that other English proficiency tests are being offered as “at home” tests. It is interesting to note that both the TOEFL iBT and IELTS were able to create and make operational “at home” versions of their tests just a few months after test centers were closed because of the COVID-19 pandemic. The fact that they had the capability to offer “at home” versions prior to the pandemic, but chose not to, suggests that their concerns about security and cheating overrode their concern about accessibility. While the DET does internal research on the issue of cheating on its test (e.g., Brem, Lake, & LaFlair, 2019), this is an issue that needs to be addressed openly, through independent, peer-reviewed, published research.

One of the innovative measures that the DET offers is that Duolingo provides the videorecorded oral interview and the writing sample to institutions that receive the scores. This serves many purposes, including as a deterrent to cheating. A test-taker knows that s/he is being videorecorded, and (at least in theory) an admissions counselor or someone at the institution could check to see once the student is on campus whether that student matches the person in the videorecorded interview. This seems like an obvious measure that all high-stakes English proficiency tests should provide for institutions that receive their scores, and it is surprising that other tests do not already provide this.

**Impact**

An important consideration for language test developers is the impact and possible washback effect that their test can have on students, teachers, and educational systems (Cheng, 2014). Millions of prospective university students take an English proficiency exam like the TOEFL iBT, IELTS, PTE Academic, and DET each year, in the hopes of gaining admission to an English-medium university. Because of the high-stakes nature of these tests, educational systems (both private and public) that prepare students for these tests inevitably will tailor their pedagogical practices to help their students succeed on these tests. The idea of “teaching to the test” is not inherently bad; indeed, if teaching to the test leads to improved pedagogical practices, and improved teaching and learning outcomes, then the test can have positive washback (Bailey, 1996). But when teaching to the test leads to a narrow focus on only passing the test (through activities like rote memorization, drills focused specifically on test tasks, the teaching of test-wiseness strategies, etc.), at the expense of teaching and learning focused at the underlying construct (English language proficiency), then that test can lead to negative washback and impact. This recognition of the importance of promoting positive washback is at least part of the reason for the move 20–30 years ago in language testing from discrete-point testing to more communicative language testing.
Most test developers specifically claim that one of the goals of their test is to have a beneficial impact on teaching and learning, and positive washback. Nowhere in the DET literature is there any mention of washback, impact, or consequences. The DET would seem to be more “preppable” than other English proficiency tests because of the discrete point nature of the test tasks. It is easy to envision test prep schools creating long lists of English words for DET test-takers to memorize in preparation for the Yes/No vocabulary tasks, which might lead to higher scores, but would do virtually nothing to improve the vocabulary knowledge of the student. Similarly, it is easy to envision pronunciation practice that focuses on parroting and the ability to pronounce correctly while reading aloud, which might lead to increased test performance, but would probably have virtually no impact on learners’ pronunciation in real-world speaking contexts.

**Discussion**

The DET was designed for “maximum accessibility” (LaFlair & Settles, 2019, p. 3), and accessibility is obviously its greatest asset. It can be taken anywhere (with a reliable internet connection), at any time, at a much cheaper cost than other English proficiency tests, in less than an hour. This focus on accessibility seems especially prescient in the COVID-19 pandemic of 2020 when test centers throughout the world were closed.

But at what cost is this focus on “maximum accessibility”? Brenzel and Settles (2017) describe how the DET was originally developed: “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” (p. 1). These are laudable goals, but nowhere is there any mention of validity, nor is there any mention of creating a test that measures communicative language ability. There is no doubt that the DET is more efficient, cheaper, and more accessible than other large-scale English proficiency tests, but if the results of the test are not valid for its intended purposes, is accessibility really all that useful?

In 2015 in this journal, I co-authored a review of the DET that was critical of its use for English-medium university admissions purposes. In July 2019, a revised version of the DET was launched. While some new tasks were introduced (https://testcenter.zendesk.com/hc/en-us/articles/360030587432-Summer-2019-Updates-to-the-Duolingo-English-Test) and the length (and cost) of the test was increased, the test is not markedly different from its previous iteration, nor did there seem to be an increased focus on assessing test-takers’ interactional competence. This seems like a missed opportunity to address the shortcomings of the test by using:

- Test tasks that require actual comprehension of spoken and written texts.
- Test tasks that tap into and measure discourse competence. As it currently stands, the only tasks that (marginally) tap into this ability are the oral interview and writing sample tasks, yet inexplicably, these tasks are not even scored, and thus do not contribute to the overall test score.
- Test tasks that use and assess academic English.
- Technology to create test tasks that are more similar to and representative of real-life language use tasks. The DET has obviously invested heavily in creating and using technology to create, deliver, and score test items, but technology can also be used to create tasks in which test-takers must interact with a text, process and comprehend it, and then respond in some way. Technology can also be used to create authentic contexts for the different language use situations.

The July 2019 revision included mostly minor changes to the test and did not address its fundamental shortcomings. Again, this revision and relaunch seem like a missed opportunity to address the many obvious shortcomings in the previous version of the test. The DET developers have
focused on utilizing technology to maximize test accessibility; why did they not utilize technology to make the test a better and more valid measure of test-takers’ interactional competence in English?

Another concern regarding the DET is the seeming lack of transparency related to the test, especially its algorithmic black box. Beyond the question of what aspect of language ability is actually being assessed with the different tasks, another question is how the scores from the different tasks contribute to the overall score. What percentage of the overall score is represented by the results of the Yes/No vocabulary tasks? What percentage do the extended writing and extended speaking tasks contribute? There is no mention of this in the DET literature. And this lack of specific information is also found in the single holistic score that a test-taker receives. Because it is unclear what aspects of communicative language ability are being assessed, it probably is not possible to give section or skills scores. Rather, there is just one holistic score that provides no useful feedback to test-takers about what aspects of English they need to improve on. With this holistic score is a “comparison table” comparing DET scores with TOEFL iBT and IELTS, and also links to two or three CEFR descriptors. The DET makes a claim about the test score interpretation, claiming that, “The scores align to the Common European Framework of Reference (CEFR), an international standard of language proficiency” (https://testcenter.zendesk.com/hc/en-us/articles/360011091711-How-do-I-interpret-my-Duolingo-English-Test-results-). For example, a DET score ranging from 90 to 115 indicates the test-taker “Can fulfill most communication goals, even on unfamiliar topics”, “Can understand the main ideas of both concrete and abstract writing.”, and “Can interact with proficient speakers fairly easily” (https://englishtest.duolingo.com/scores). This stretches credulity, as there is no meaningful empirical evidence on which to make this score interpretation claim.

With language test validation, be it a more traditional construct validation approach, or an argument-based approach, it is necessary to have both a strong theoretical rationale and robust empirical evidence supporting a test’s use for a particular purpose. For uses that involve high stakes decisions (like university admissions), the need for theoretical and empirical support is even stronger. If a language test employs test tasks that are similar to the real-world language tasks in the domain of interest, it is easier to make that theoretical argument. As described above, the characteristics of the DET test tasks have little in common with the characteristics of language tasks that university students are commonly expected to perform. Test tasks do not need to be “authentic”, of course, but if they are not, then the empirical evidence in support of the validity argument needs to be that much more sound and robust. However, the research supporting the validation of the DET for English-medium universities is woefully short of this standard. It seems that virtually all of the research that has been done was conducted by or commissioned by DET, rather than independent and peer-reviewed research. Under the “Research” tab on the DET website are listed a number of DET research studies, but only two (Maris, 2020; Settles et al., 2020) seem to be focused on the revised version (July 2019) of the test, and neither one is a true validation study. The DET technical manual concludes with a section titled “Relationship with Other Tests”, which apparently refers to concurrent validity, and describes how starting in August 2018 test-takers were asked to share recent scores they had received on the TOEFL iBT and IELTS. These scores were used to correlate DET scores with TOEFL iBT and IELTS, and also to create concordance tables. Based on this internal, unpublished research involving test-takers’ self-reported scores, LaFlair and Settles (2019) conclude that, “The research reported here illustrates evidence for the validity of the interpretations and uses of the Duolingo English Test” (p. 19).

On the DET website and in the DET literature (e.g., Brenzel & Settles, 2017; LaFlair & Settles, 2019), it is stated repeatedly that DET scores are used for university admissions purposes. The DET website has a section devoted to “Admissions” (https://englishtest.duolingo.com/edu), boasts that DET scores are “accepted by hundreds of universities worldwide”, and list over a thousand institutions that accept DET scores as part of the admissions process. But to my knowledge, it is not stated on the website or in any of the DET research literature that the DET was actually developed or validated for this use. In order to make any claims about the validity of a test, the purpose needs to be explicitly delineated, because a test is only valid for its intended use (Chapelle & Voss, 2014).
There is very little empirical evidence supporting its use for university admissions purposes, and this makes it incredibly difficult to present a convincing validation argument supporting the use of DET scores for admissions purposes at English-medium universities.

**Conclusion**

The (revised) Duolingo English Test was designed for maximal accessibility. It can be taken anywhere, in less time and for less money, and results are delivered to test-takers in less than 48 hours. This focus on accessibility was prescient, as witnessed by its competitors (including TOEFL iBT and IELTS) not being able to administer their tests in many markets because of closures to test centers because of the COVID-19 crisis of 2020. And even though TOEFL iBT and IELTS created “at home” versions of their own tests soon after, the DET continues to be more accessible than its competitors in that it takes less time to take, costs less, and the results are delivered to test-takers more quickly. In addition, the DET shares a writing sample and a videorecorded interview of the test-taker with the institutions that receive the scores from Duolingo. This is a useful service (and one that should be copied by the DET’s competitors).

But in its current formulation, I cannot recommend that DET scores be used for university admissions purposes. The DET was revised in July 2019, but this revision did almost nothing to address its many weaknesses: its reliance on inauthentic test tasks that have little in common with the types of language uses required in university contexts; the black box nature of its scoring procedures; its failure to assess discourse-level competence, interactional competence, and academic language; its susceptibility to cheating; its possible negative washback effects on learners and learning systems; and an almost total lack of independent, empirical, peer-reviewed research validating the use of DET scores for university admissions purposes.

**Disclosure statement**

No potential conflict of interest was reported by the author.

**ORCID**

Elvis Wagner http://orcid.org/0000-0003-2332-3323

**References**


Brem, C., Lake, J., & LaFlair, G. (2019). *A statistical index of cheating (and rule breaking) in a high-stakes computer adaptive language assessment*. Paper presentation at the annual meeting of the Midwest Association of Language Testers (MwALT), Bloomington, USA.


