Routledge
Taylor & Francis Group

Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: http://www.tandfonline.com/loi/hlaq20

The Duolingo English Test

Elvis Wagner & Antony John Kunnan

To cite this article: Elvis Wagner & Antony John Kunnan (2015) The Duolingo English Test, Language Assessment Quarterly, 12:3, 320-331, DOI: 10.1080/15434303.2015.1061530

To link to this article: http://dx.doi.org/10.1080/15434303.2015.1061530

Published online: 07 Aug 2015.

Language Assessment Quarterly, 12: 320–331, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1543-4303 print / 1543-4311 online
DOI: 10.1080/15434303.2015.1061530

TEST REVIEW

The Duolingo English Test

Elvis Wagner
Temple University, Philadelphia, Pennsylvania

Antony John Kunnan
Nanyang Technological University, Singapore

INTRODUCTION

This article provides a critical conceptual review of the Duolingo English test (DET) and explore possible consequences of its use for university admissions purposes. Because the DET is very new, the only published research article supporting its claims is Ye (2014); but this study was commissioned by Duolingo and is not peer reviewed. In addition, there is very little publicly available material about the development of the test or the model of language ability that it purports to be assessing. Nevertheless, we decided to critique the test because of the high-profile nature of the test, and its potential to have a disruptive influence (positive and negative) on the language-testing industry.[1]

THE DUOLINGO ENGLISH TEST

The DET was created by the developers of the Duolingo Language Learning program, a free, Web- or application-based language-learning program (https://www.duolingo.com/). The DET purports

[1] The authors are independent reviewers; they were not funded for this review by Duolingo English Test or any of its competitors. The first author is currently conducting a TOEFL grant-funded study and the second author is currently conducting a GEPT grant-funded study.

Correspondence should be sent to Elvis Wagner, Teaching and Learning Department, Temple University, 447 Ritter Hall, 1301 Cecil B. Moore Avenue, Philadelphia, PA 19122. E-mail: elviswag@temple.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hlaq.

to assess “proficiency in daily English” (Ye, 2014, p. 4)[2] of L2 learners in reading, writing, listening, and speaking. According to Burr (personal communication, February 9, 2015), it is not a test of proficiency in academic English; rather, it is a measure of general English skills and usage. The DET’s website claims that the test is “…scientifically designed to provide a precise and accurate assessment of real world language ability” (https://testcenter.duolingo.com/faq).

It is a computer-adaptive test, meaning that an algorithm determines the items that each test taker will be presented with, depending on how he/she answered previous items. Because it is computer adaptive, the number of items that each test taker must answer may differ; consequently, the length of the test varies for each test taker. The average test takes approximately 16 minutes, and the maximum length of the test is capped at 20 minutes.

The test was developed on the basis of a data-driven model of language learning.[3] A corpus of texts that had been classified at particular Common European Framework of References for Languages (CEFR) levels were mapped on to a numerical scale, and then by using machine learning and natural language processing, a number of models were created to map other texts to be used with the DET onto the CEFR. In other words, the CEFR level of previously unrated texts are rated by using the models that have been created through the use of machine learning and natural language-processing models, and these texts are then used in determining the difficulty level for use in the computer adaptive test (Burr, personal communication, February 9, 2015). All the tasks and items are computer scoreable.

CURRENT USES

The website for the DET (https://testcenter.duolingo.com) urges potential test takers to “Certify your English proficiency” and “Upgrade your resume” by taking the test and advertises that it provides “affordable and convenient language certification,” although what the “certification” is used for is not specified. Prominent on the test homepage is a link that says “Universities and institutions: Contact us to accept Duolingo English Test scores.” According to Burr (personal communication, February 9, 2015), the test is currently used by some non-U.S. English-medium universities for admissions purposes. It is anticipated that more universities, including U.S. universities, will use the test for admissions purposes, and there is currently a validation research project underway involving admissions data from a number of U.S. universities. The DET website states, “We are the most accurate language learning certificate in the world. Universities, organizations and companies around the world are beginning to accept the Test Center certificate. We’ll keep you updated as more institutions sign on” (https://testcenter.duolingo.com/faq). Thus, the ultimate purpose of the test seems to be to use it as an assessment of the English language proficiency of non-English-speaking international students seeking admission to English-medium North American universities, similar to the TOEFL iBT, the IELTS, and the Pearson Test of English Academic.

[2] Because there is very little published material about the goals and development of the test, much of the information here is based on Ye (2014) and from the test’s webpage: https://testcenter.duolingo.com/. The screen shots illustrating the different task types are taken from Ye (2014).

[3] Again, because of the lack of publicly available information about the test’s development process, much of the information here is based on a phone conversation with one of the test developers (Burr, personal communication, February 9, 2015).

TEST TASKS

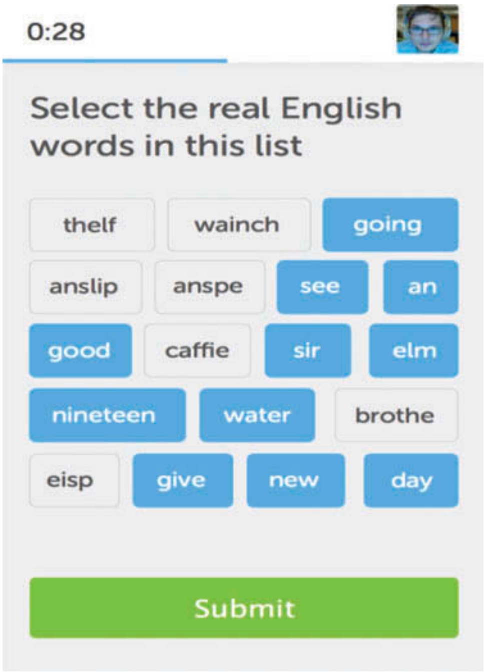
There are four separate tasks on the DET: a vocabulary task, a listening and transcription task, a sentence completion task, and a speaking task. As indicated earlier, because the test is computer adaptive, the order and number of times each task appears may differ for each test taker on the basis of his or her responses.

Vocabulary

In the vocabulary task presented in Figure 1, the test taker sees a number of words (approximately 18) in boxes on the screen. The test taker must select the words that are actual English words. The test taker has one minute to complete the task.

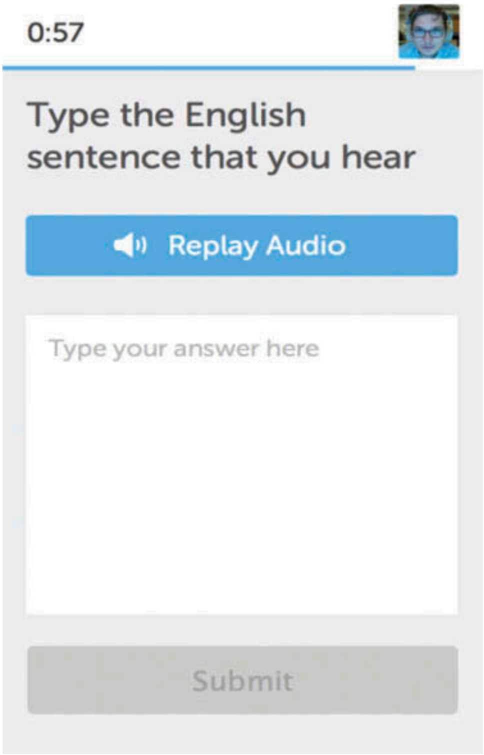
Listening and Transcription

This task presented in Figure 2 purports to be assessing test takers' listening ability. A sentence is presented aurally to test takers, and they must type the sentence that they heard. This task is essentially a dictation task. Test takers have one minute to complete the task and they can hear the audio text up to three times (by clicking on the "Replay Audio" button, as shown in Figure 2).



Select the real English words in this list

FIGURE 1 Vocabulary task.



0:57

Type the English sentence that you hear

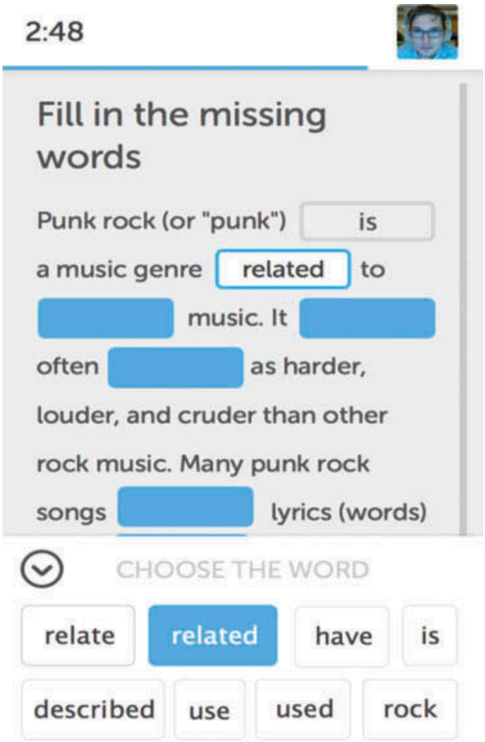
FIGURE 2 Listening and transcription task.

Sentence Completion

This task, presented in Figure 3, is labeled as a sentence completion task. Here, the test taker is presented with a short text composed of several sentences. There are five deleted spaces in the task, and the test taker must choose the appropriate word from a list of eight words or morphemes by clicking on the blank space to fill the deleted space. For each of the five deleted spaces, the same eight words or morphemes are given. This task is similar to a selected-response or rational cloze task. The test taker has three minutes to complete the task.

Speaking

The fourth task presented in Figure 4 is a speaking task. Here, test takers see a written sentence and are instructed to "Speak this sentence." Test takers click on a microphone icon and read aloud the written sentence, and they then click on the microphone icon to stop. The test taker has one minute to complete the task.



2:48

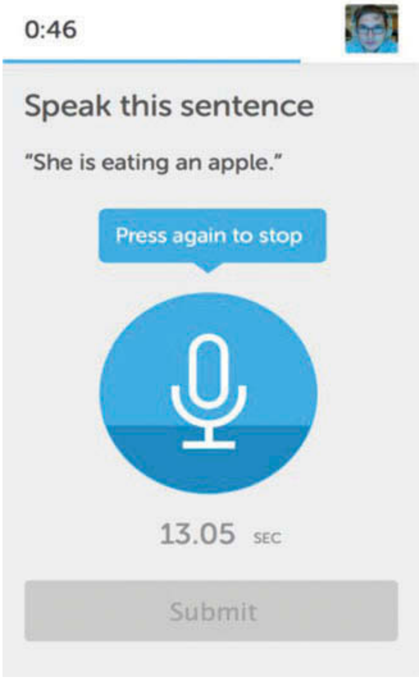
Fill in the missing words

Punk rock (or "punk") is
a music genre related to
music. It
often as harder,
louder, and cruder than other
rock music. Many punk rock
songs lyrics (words)

FIGURE 3 Sentence completion task.

ADMINISTRATION

The test is administered by computer, mobile device, or smart phone, using the device’s screen, camera, keyboard, speakers, and microphone. The test taker is urged to take some of the sample questions available before actually taking the test, to become familiar with the test format, and to become comfortable using the camera, microphone, and speakers on the computer. After accessing the site and registering, the test taker is informed that a supervisor will review a recording of the test to prevent cheating. The test taker must display his or her government-issued identification, and a picture of it is taken by the computer’s camera. The test taker is warned not to speak unless instructed, not to look away from the screen, not to have his or her face leave the camera preview, and not to use other devices. These measures are taken to ensure that the test taker does not receive any assistance. After the verification process is complete and $20 is paid, the test begins. Test takers complete four different test tasks (described above) of differing difficulty levels, based on their responses to previous tasks. The number of tasks they must complete may differ.



0:46

Speak this sentence

"She is eating an apple."

FIGURE 4 Speaking task.

SCORING PROCESS

Because the test is computer adaptive, the scoring is completely automated and immediate. For the vocabulary task, the listening and transcription task, and the sentence completion task, the test taker selects a response, and scoring is based on correct or incorrect responses. For the speaking task, in which the test taker speaks a sentence, the computer scores the response using a proprietary algorithm. There is no information available in the public domain about how the test is being scored, but presumably the algorithm was developed through machine learning by analyzing a corpus of spoken texts, and a number of measures, including speech rate, length of pauses, phoneme accuracy, and intonation, are all measured quantitatively to derive the score. The results are provided to the test taker within 48 hours. The test taker receives a score from 1 to 10, with a short description of what the score means and what a test taker at a particular level can do in English.

TEST QUALITIES

Any critique of a language test has to be based on qualities or properties that a test contains and the claims stated explicitly (on their website or in published research) or implicitly (inferred from the test's purpose and score use) about its capabilities. There are two test critique approaches used currently. In the first approach, external standards for tests and testing practice have been formulated with detailed standards by educational professionals (of the AERA, APA, NCME Standards, 1999, 2014), the International Language Testing Association (ILTA, Code of Ethics and Practice, 2000), and individual researchers (Bachman & Palmer, 1996, 2010), and Kunnan (2014). These documents identify many aspects that include validity or meaningfulness, reliability of scores, absence of bias, appropriate access and administration, and beneficial consequences. The second approach proposed by Kane (2012) prefers the test developer to offer their own interpretive argument, which then is investigated through a validity argument by the test developer or an independent researcher. We chose to use the former approach; therefore, we critiqued this test from the following aspects: validity or meaningfulness, reliability or consistency, absence of bias, access, and consequences.

Validity or Meaningfulness

As stated earlier, the DET will potentially be used as a measure of international applicants' English proficiency for admission to English-medium universities in North America. Yet, according to Burr (personal communication, February 9, 2015), the test purports to be a test of general English proficiency and skills and not a test of academic English ability. This gap between the language (and task) characteristics in the target language use domain (of academic tasks similar to tasks in university courses) and the characteristics of the language and tasks used in the test is a fundamental shortcoming of the test.

A second fundamental shortcoming of the DET is that it measures only a very limited construct of English proficiency. Because of the need to have a test that can use fully automated scoring, the types of tasks used on the test are severely constrained (transcribing, reading aloud, vocabulary, lexico-grammatical cloze), resulting in a very narrow operationalization of the construct of English proficiency. This problem is often referred to in the literature as under-representation of the construct.

Using tasks that can only be scored automatically also precludes the assessment of test takers' ability to use English *interactively*, with other human language users. There is a seemingly universal consensus in the field of second language acquisition that the overarching goal for L2 learning and instruction should be communicative competence and performance—to prepare learners to be able to use the target language to communicate with other speakers of that language. Indeed, this theory of communicative competence has been the "central doctrine" (Leung, 2005, p. 124) in English language learning and teaching for over 40 years. Yet the DET does not seem to tap into test takers' communicative competence.

The DET also lacks any actual *language production* by test takers. Instead, all of the tasks require the test takers to identify actual English words, transcribe a spoken sentence, read aloud a written sentence, or choose the correct morpheme or word to fill in the blank. This indirect assessment of speaking and writing ability presents major threats to the validity of the test's use for university admissions purposes.

The test focuses almost exclusively on word and sentential level ability and very little at the discourse level. The vocabulary task is word level only. The listening and speaking tasks are at the sentence level and not beyond. The "sentence completion" task (which is actually a cloze task requiring lexical and grammatical ability, and some reading ability) is the only task that goes beyond the sentential level, and even this task involves a text that is at maximum two or three sentences in length. This lack of any language use at the discourse level presents real threats to the validity of using these test results as a measure of English proficiency, especially as a measure of a test taker's ability to perform in an English-medium college or university. In addition, because of the lack of discourse level texts, the DET does not assess test takers' pragmatic or sociolinguistic competence in any way (Purpura, 2004).

Finally, the test tasks used are problematic for multiple reasons. With the vocabulary task, the test taker must differentiate between the "real English" words and the nonsense words. Whether this ability to recognize actual English words is a relevant ability to be assessing is unclear, but this type of "spot the real English word" vocabulary task certainly does not assess the test takers' productive knowledge of the meaning of the words or the test takers' ability to use that word in writing or speech (Laufer, 1998; Nation, 2001).

Similarly, the listening task used in the test has numerous shortcomings. The listening task does not measure comprehension. Rather, it measures the test takers' ability to transcribe a sentence. Indeed, this task requiring listeners to transcribe word for word is very little like real-world listening, which requires the listener to segment the spoken input, assign semantic and syntactic meaning to the decoded input, compare this information with the listener's background/contextual knowledge, knowledge of the co-text, and sociolinguistic and pragmatic knowledge, and make numerous inferences about the speaker's meaning. In addition, in many real-world listening contexts, there would be an interactional component, in which the listener is also trying to formulate an appropriate response to the spoken text. Furthermore, the spoken texts used on the listening task are only at the sentence level (this lack of discourse level texts has already been critiqued above), and these spoken texts are spoken slowly with explicit enunciation, and therefore, lack the characteristics of connected speech that are prevalent in most real-world, unplanned spoken language (Wagner, 2014). It should also be noted that this listening task is the only instance in the test where the test taker is required to actually write anything, and this is only transcribing a spoken sentence. For university admissions purposes, writing ability would seem to be a critical component to be assessed, yet the DET does not measure writing ability of any kind (much less academic writing ability).

The speaking task requires test takers to read a sentence aloud. Speaking is assessed only indirectly and does not assess the test takers' interactional or sociolinguistic competence. Indeed, it would seem that only pronunciation is assessed (and only a very narrow operationalization of pronunciation—the ability to pronounce correctly while reading a sentence aloud). This issue of pronunciation being the sole component of speaking competence is even more troubling, considering that there is no information about which aspects of pronunciation are being assessed. The obvious concern here relates to the issue of accentedness and intelligibility. Accentedness and intelligibility are related yet separate constructs (e.g., Derwing & Munro, 2009; Isaacs, 2014). That is, a person can speak English with a "foreign" accent, yet still be highly intelligible to listeners. But there is no information available about the issue of accentedness or dialect with the DET. Similarly, there is a general consensus in the field of theoretical linguistics that no one variety, dialect, or accent of English is superior to another (e.g., Jenkins, 2006; Lippi-Green, 2011), yet it is unclear if the scoring algorithm for the speaking test takes this into consideration. If a test taker speaks with a "non-standard" dialect, will he or she be penalized? These are serious issues that need to be addressed by the test developers. Nevertheless, however serious these concerns are, they are minor in comparison to the much larger problem that pronunciation ability while reading aloud is a very poor proxy measure for the type of speaking ability needed to succeed in an English-medium university.

The sentence completion task (reading cloze) seems to be measuring some mixture of vocabulary knowledge, morphological knowledge, grammatical knowledge, discourse knowledge, and reading comprehension. There is some evidence in the literature that the scores from these types of cloze tasks correlate with overall reading comprehension (Bachman, 1982; Alderson, 2000). But cloze tasks are at best only an indirect measure of reading comprehension, and the cloze task on the DET is not even at the paragraph level, although university students are expected to be able to read and comprehend extensive written texts.

In summary, in validity or meaningfulness, the DET has a number of fundamental flaws and shortcomings that make its use for English-medium university admission unsupportable: The test does not measure a test taker's ability to interact in English; the test focuses almost exclusively on word and sentential level language, with no assessment of test takers' ability to understand or produce extended texts; the test focuses on test takers' receptive knowledge at the total exclusion of assessing the test takers' productive knowledge; the language and tasks the test takers encounter on the test are very unlike the types of language and language tasks the test taker would encounter, process, and produce outside of the test context and in the target language use domain; and the test does not assess test takers' ability to use and comprehend academic English, even though the apparent use of the results from the test are for university admission purposes.

There is an almost total lack of published research on the use of the DET for university admissions purposes. In fact, the only research study that has been published examining the validity of the test is Ye (2014), who found that DET scores correlated substantially (r = .67) with composite TOEFL iBT scores, and moderately with iBT section scores (Reading, r = .45; Listening, r = .52, Speaking, r = .56; Writing, r = .56). This type of criterion-related validation study is insufficient as an assessment argument for its use for university admissions purposes. While the DET is currently the focus of a larger research study examining its use for university admissions purposes (Burr, personal communication, February 9, 2015), it seems unlikely that a credible validation argument can be built for the test, at least in its current form.

Reliability or Consistency

Because the test is entirely computer scored, reliability would seem to be a strong point of the test. No expert rating or judgment is required for any of the scoring, and thus, difficulties with reliability typical of human scoring are not a problem for the DET. Ye (2014) reported a test-retest reliability coefficient of .79, which is reasonably high, especially considering that it is a computer-adaptive test. However, there is no report available of the internal consistency reliability of the tasks, nor is there any report on the agreement of the automated scoring with human judgments. These reliability statistics need to be studied and made available to the public by the test developers.

Absence of Bias

There is no published research that has investigated the issue of bias with the DET, although there are plans to research this issue in the future (Burr, personal communication, February 9, 2015). An important consideration for future research is how the performance of test takers with little or no familiarity or expertise in computer use might be affected by the need to use computers or smartphones when taking the test. In addition, it would be necessary to examine the content of the tasks (topics) and test performance for possible bias against types of test takers (by gender, race and ethnicity, socioeconomic status, and international status; see Kunnan, 2007, for examples of such research).

Access

The DET's low-cost barriers and ease of accessibility due to not having to travel to a test center would seem to be the test's most positive quality. However, for potential test takers who do not have access to computers or smart phones, or to the Internet, taking the test is not possible. In addition, the test will not be accessible for those potential test takers without computer familiarity.

The issue of access also applies to test takers with disabilities. To date, no research has been conducted regarding how test takers with disabilities might perform on the test, nor have there been any studies examining possible accommodations to test takers with disabilities, although this type of research is planned for the future (Burr, personal communication, February 9, 2015). In the United States this would be in violation of the Americans with Disabilities Act of 1990.

Consequences

The DET test has the possibility of having a profound impact on future test stakeholders, including test takers, universities and colleges, and even test developers. The low cost and convenience of the DET is hugely beneficial to test takers and is also attractive to universities and colleges, who want to keep barriers as low as possible to potential applicants. But wide acceptance of the DET for university admissions purposes also has the possibility of having profound negative consequences for test stakeholders. As has been argued throughout this critique, the language and the language tasks that are used in the test are fundamentally different from the language and language tasks found in the higher-education target language use domain. Because of these fundamental differences, a very strong validity argument needs to be presented by the test developers demonstrating the appropriateness of the test's use for university admissions purposes. Yet this validity argument has not been made.

Many researchers have argued that a test should be beneficial to the community it serves (see Bachman & Palmer, 2010; Kunnan, 2014). Specifically, there should be tangible benefits to a community along aspects we have discussed such as validity, link to learning, reliability, access, and consequences. In the DET it is unclear how beneficial the test is going to be to the community except in terms of access (cost, ease of use, and test taking time). But surely, the benefits have to be more than just these elements. Furthermore, public reasoning or justification of a test's claims by its developers is argued to be a necessary component of developing, launching (including charging fees), and reporting scores and providing descriptors of ability (see Kunnan, 2014 for a detailed argument regarding this point).

Potential for Cheating

Developers of standardized tests of English proficiency face increasing instances of cheating. Because these tests are high stakes, test takers (and the test preparation industry) devote considerable resources to gaming the testing system, as recent cheating scandals involving English proficiency tests demonstrate (see http://chronicle.com/article/English-Testing-Companies-Vie/123671/). Because there is no human proctor physically present during the test taking, the possibility for cheating seems enhanced with the DET. While the test has a number of procedures built into it as safeguards against cheating, it is unclear how robust and how effective these procedures are, and it seems likely that test takers will devise ways to overcome them. An obvious way to cheat on the test would be for a person to ask (or pay) someone else more proficient in English to take the test for him, using the first person's identification documents or a falsified identification document.

CONCLUSION

In summary, at the time of writing this critique, the DET seems woefully inadequate as a measure of a test taker's academic English proficiency or for high-stakes university admissions purposes. We ask this fundamental question: Is the DET really assessing a test taker's ability to understand and communicate with other speakers of English in a university setting? There is virtually no research showing that it is.

The test seems to be a case of "the tail wagging the dog," in that the DET's reliance on short, computer-scored test tasks has resulted in a test that does not assess the test takers' communicative competence. Indeed, the test tasks that are used hearken back to the 1950s, when audiolingualism was the dominant theory in language learning. As with audiolingualism teaching techniques, the DET has many tasks that focus on pronunciation, grammar, and receptive vocabulary knowledge but do not involve, require, or assess test takers' communicative or interactional competence. The test takers do not have to compose and produce any language (speaking or writing), nor do they have to comprehend the meaning of spoken or written input (reading and listening comprehension). The DET test seems to completely ignore decades of accumulated knowledge and research about how languages are best learned and how language proficiency can best be assessed.

Finally, with no public reasoning or justification in support of the test's claims in public forums or through published journal articles or research reports in the public domain on their website (except the inadequate report from Ye, 2014), we urge test score users (e.g., North American English-medium university admission officers) to be extremely cautious with the use of test takers' scores from the DET.

REFERENCES

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, *16*, 61–70.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

Derwing, T., & Munro, M. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*, 1–15.

Isaacs, T. (2014). Assessing pronunciation. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 1, pp. 140–155). Malden, MA: Wiley.

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELT Journal*, *60*, 42–50. doi:10.1093/elt/cci080

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*, 3–17.

Kunnan, A. J. (2007). Test fairness, test bias & DIF. *Language Assessment Quarterly*, *4*, 109–112. doi:10.1080/15434300701375865

Kunnan, A. J. (2014). Fairness and justice in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 3, pp. 1098–1114). Malden, MA: Wiley.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Journal of Applied Linguistics*, *19*, 255–271. doi:10.1093/applin/19.2.255

Leung, C. (2005). Convivial communication: Recontextualizing communicative competence. *International Journal of Applied Linguistics*, *15*, 119–144. doi:10.1111/ijal.2005.15.issue-2

Lippi-Green, R. (2011). *English with an accent* (2nd ed.). London, UK: Routledge.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. New York, NY: Cambridge University Press.

Purpura, J. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

Wagner, E. (2014). Using unscripted spoken texts in the teaching of second language listening. *TESOL Journal*, *5*, 288–311. doi:10.1002/tesj.2014.5.issue-2

Ye, F. (2014). Validity, reliability, and concordance of the Duolingo English Test. Retrieved from https://s3.amazonaws.com/duolingo-certifications-data/CorrelationStudy.pdf