ISSN 2515-1703
May 2022

IELTS Partnership Research Papers

Comparison of IELTS Academic and Duolingo English Test

Sara T. Cushing and Haoshan Ren, Georgia State University

IELTS™
BRITISH COUNCIL idp Cambridge Assessment English

Foreword

This report provides an in-depth comparison between IELTS Academic and the Duolingo English Test (DET), based on a review of publicly available documentation and published scholarship on each test. We follow the analytical framework found in Taylor and Chan (2015), who employed and expanded on the socio-cognitive framework (SCF) for test validation introduced by Weir (2005). This paper is framed by the six components of the SCF: test taker characteristics, cognitive validity, context validity, scoring, consequences, and criterion-related validity.

In terms of test taker characteristics, our analysis of published demographic data suggests that the population of test-takers for each test are approximately equivalent in overall proficiency. While IELTS Academic is specifically designed for use in educational settings, DET was originally designed as a general proficiency test. However, some recent Duolingo publications have stated that its main purpose is for admissions decisions.

To compare cognitive and context validity of the two tests, our analysis focuses on the four main language skills (reading, listening, speaking, and writing) and the specific test tasks targeting each skill. For all four skills, IELTS tasks elicit a wider range of cognitive processes than the DET tasks, and the DET items are generally less oriented to academic skills required in higher education contexts. In terms of scoring validity, despite large differences in the way scores are calculated, both tests appear to be scored reliably and to demonstrate internal consistency, and both testing organizations seem to have in place sufficient procedures for monitoring test performance. Our analysis of criterion-related validity suggests that there is a relationship between scores on the two tests; however, this relationship needs to be interpreted with caution. In particular, we were unable to find any publicly available information about how DET mapped its scores onto the CEFR. Finally, by analyzing available online discussions about the two tests, we discuss their consequential validity. Given that many test takers are focused on getting the highest possible scores on tests, our analysis suggests that the test preparation strategies recommended for IELTS may be more applicable to future academic work than those for DET.

In conclusion, we found that, compared to IELTS, DET test tasks under-represent the construct of academic language proficiency as it is commonly understood, i.e., the ability to speak, listen, read, and write in academic contexts. Most of the DET test tasks are heavily weighted towards vocabulary knowledge and syntactic parsing rather than comprehension or production of extended discourse. Scores on the two tests are correlated, which might suggest that DET could be a reasonable substitute for IELTS, given its accessibility and low cost. However, even though knowledge of lexis and grammar are essential enabling skills for higher-order cognitive skills, a test that focuses exclusively on these lower-level skills is probably more useful for making broad distinctions between low, intermediate, and high proficiency learners rather than for informing high-stakes decisions such as university admissions.

Comparison of IELTS Academic and Duolingo English Test

Funding

This research was funded by the British Council and supported by the IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia.

Publishing details

Published by the IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia © 2021.

How to cite this paper

Cushing, S. T., & Ren, H. (2022). Comparison of IELTS Academic and Duolingo English Test. *IELTS Partnership Research Papers, 2021/1*. IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia. Available at https://www.ielts.org/for-researchers/research-reports

Authors' biodata

Sara T. Cushing

Sara Cushing is a Professor of Applied Linguistics at Georgia State University. She has published research in the areas of assessment, second language writing, and teacher education and has been invited to speak and conduct workshops on second language writing assessment throughout the world, most recently in Vietnam, Colombia, Thailand, and Norway. Her most recent publications focus on the intersections between corpus linguistics and assessment.

Haoshan Ren

Haoshan (Sally) Ren is a Ph.D. candidate in the Department of Applied Linguistics and English as a Second Language at Georgia State University. Her research interests include language assessment, corpus linguistics, and sociolinguistics, using both qualitative and quantitative methods.

Contents

Tables and Figures

I. Introduction

For decades, universities and other educational institutions have relied on large-scale English proficiency tests to assess whether prospective students coming from other countries have sufficient English proficiency to meet the listening, speaking, reading, and writing demands of an academic curriculum.

In a high-stakes testing situation such as academic admissions, test developers must balance competing concerns in offering a valid and secure test to a heterogenous international population of test takers. Most assessment professionals agree that it is important to sample the relevant domain of target language use widely and to develop test tasks that are broadly representative of authentic language tasks and invoke the relevant language skills. Examples of academic test tasks include answering comprehension questions based on short lectures or excerpts from textbooks, or writing formal academic essays that present and support a point of view. These can be distinguished from more general language proficiency test tasks by their focus on oral and written genres that are typical of a university setting.

At the same time, the desire to test prospective students' ability to use language in academic contexts, as opposed to testing their knowledge of specific linguistic forms, inevitably requires sufficient time for test takers to demonstrate comprehension of relatively complex materials and produce extended discourse in speaking and writing, so such tests frequently last more than two hours, leading to fatigue on the part of test takers. Furthermore, speaking and writing tasks must be evaluated, typically by human raters, which adds to both the cost and turnaround time between test and score reporting. One of the most frequently expressed critiques of large-scale tests is that they put undue burdens on students in terms of time and money (see, for example, Pearson, 2019).

The IELTS Academic test, which is one of the two most widely used proficiency tests worldwide (the other being TOEFL iBT), has a long and well-documented history of research and development justifying its use for academic admissions, though it is certainly not beyond criticism (e.g., Pearson, 2019; Pilcher & Richards, 2017). As with any high-stakes, large-scale test, the resources to develop and pilot new test items and forms, maintain test security, and provide face-to-face interviews, add to the cost of the tests discussed above.

Advances in technology such as machine learning (ML) and natural language processing (NLP) have opened up possibilities for newer tests that promise to provide useful information about prospective students' language proficiency at a lower cost and with a shorter timeline. One such test is the Duolingo English Test (DET), which was first developed in 2016. Duolingo's solution to the financial demands of tests like IELTS is to use "test item formats that can be automatically created, graded, and psychometrically analyzed using ML/NLP techniques. This solves the 'cold start' problem in language test development, by relaxing manual item creation requirements and alleviating the need for human pilot testing altogether" (Settles, LaFlair & Hagiwara, 2020). In 2020, during the COVID-19 pandemic, DET gained prominence as a temporary alternative to IELTS for university admissions when many testing centers were forced to close, due to its accessibility as a remote-proctored online test. Now, however, many institutions have questions about the usefulness of DET test scores relative to IELTS for making admissions decisions for academic study and wonder what criteria to use when determining what tests to accept as evidence of English language proficiency.

The purpose of this paper is to provide an in-depth comparison between IELTS Academic and the DET, particularly in terms of test content and important aspects of test validity. Our analysis is based on a review of publicly available documentation on each test, along with published scholarship about the two tests. For comparing test content, we relied on

sample test questions available from the official websites of the two organizations (IELTS.org and https://www.englishtest.duolingo.com/home, respectively). Since the number of published test items for DET is quite small, and there is relatively little published research on the test, the first author of this report also took two sample tests, capturing screen shots of the items presented during each practice test. Following the introduction of the new Interactive Reading task, the second author took a practice test to obtain a sample passage for this task type. Like the operational DET, items in a sample test are delivered adaptively; that is, if one question is answered incorrectly, a relatively easier item is presented next (see further discussion below). In this way it was possible to gain access to items at a variety of difficulty levels. In the first test, an attempt was made to simulate responses that would be made by a less proficient English language user, and in the second, an attempt was made to answer all items as accurately as possible.

It should be noted here that the bulk of our analysis was conducted in 2021. However, in March 2022, Duolingo published a revised technical manual and two additional reports updating the reading and writing portions of the test. Where feasible, we have incorporated these updates into this document.

In analyzing the two tests, we follow the analytical framework found in Taylor and Chan (2015), who compared several English language tests to investigate their comparability to IELTS in terms of their suitability for certifying the English language proficiency of doctors applying to work in the United Kingdom. Taylor and Chan provide in-depth analyses of the four skill areas (reading, listening, speaking, and writing) for each of the tests, employing the socio-cognitive framework (SCF) for test validation introduced by Weir (2005) and expanded by scholars at Centre for Research in English Language Learning and Assessment (CRELLA) over the past several years (see, for example, Chalhoub-Deville & O'Sullivan, 2021).

As Taylor and Chan note, this framework provides "a coherent and accessible methodology for test development and validation research" (p. 27) that can be used to analyze language tests, particularly in terms of identifying aspects of the test where the construct is under-represented or includes construct-irrelevant features. The SCF consists of the following components, each of which has a set of guiding questions that can be useful in critically evaluating tests (see https://www.beds.ac.uk/crella/about/socio-cognitive-framework/):

> **Test taker characteristics:** Who takes the test? Where and how do they need to use the language?
>
> **Cognitive validity:** Do test takers engage the same cognitive processes when using language for the test as in real life?
>
> **Context validity:** How do the tasks on the test represent the ways in which test takers will use the language?
>
> **Scoring:** Do the scores reflect the importance of target skills? Are the scores reliable?
>
> **Consequences:** How does the use of the test affect teaching and learning? Does use of the test benefit society?
>
> **Criterion-related validity:** Do scores on the test match scores on other tests of the same abilities? How well does the test predict performance in real life?

We go into more detail about the components of the SCF in the relevant sections of the report, which is organized as follows. First, we present an overview of the two tests. This is followed by a discussion of test taker characteristics for both tests. Next, we look at both cognitive and context validity in terms of the four main language skills: reading, listening, writing, and speaking. We then consider scoring, consequences, and criterion-related validity in the final section of the report.

II. Overview of the two tests

International English Language Testing System (IELTS)

The International English Language Testing System (IELTS) is a globally recognized English proficiency test for non-native English speakers who intend to work, study, or migrate to a country where English is the predominant language. It measures and reports on the four main language skills – listening, reading, speaking, and writing. IELTS scores for the whole test or, in some cases, scores on individual subskills, are accepted as evidence for English proficiency in a variety of industries, academic institutions, and immigration bodies, especially in Australia, Canada, New Zealand, and the UK. Jointly owned by The British Council, IDP: IELTS Australia, and Cambridge Assessment English, IELTS has been in operation for four decades. Along the way, IELTS has developed three modes of delivery: paper-based, computer-delivered, and online (IELTS Indicator). The task types are the same in all three modes.

The paper-based IELTS test has been the primary delivery mode since the launch of the test in 1989. Test-takers take the Reading, Listening, and Writing sections of the test in one sitting at a designated testing site, and participate in a separate face-to-face session with a certified IELTS examiner for the Speaking portion of the test. Similarly, the computer-delivered IELTS test requires test-takers to take the Reading, Listening, and Writing sections in official IELTS testing centers, and take the Speaking test face-to-face separately with a certified examiner. The test report, content, timing, and structure are the same for both the paper-based and the computer-delivered test, except that the computer-delivered test has a slightly shorter time limit for its Listening section, taking into consideration that test takers do not need to manually transfer their answers to an answer sheet. IELTS Indicator is an IELTS online test developed to cope with the lockdown of IELTS testing centers during COVID-19. Test takers can take the online exam at home, and the test is designed with the same structure and content as the paper- and computer-based tests. According to the IELTS official website, the IELTS Indicator only provides an indicative score, which is accepted by a limited number of institutions.

All IELTS test scores are converted to band scores from 0–9, with 9 indicating the test taker has an expert level of the operational command of English, and 0 being assigned to test-takers who did not attempt the test. The paper-based and computer-delivered tests have two alternative versions: IELTS General Training and IELTS Academic, assessing language use for different purposes. The IELTS Indicator is only designed for academic purposes.

The two modules differ in their Reading and Writing sections, while the Listening and Speaking sections are the same. In this report we are focusing on IELTS Academic rather than IELTS General Training, since our focus is on tests for university admission.

The Listening section consists of four recorded monologues and conversations, and question types include short answer, form completion, multiple choice, matching, sentence completion, plan/map/diagram labeling and note completion.

The Academic Reading section has texts taken from books, journals, magazines and newspapers. The texts are presented to test takers at the same time as the questions, which include matching headings, multiple choice (more than one answer), identifying information, note completion, reading summary completion (selecting words from the text, or selecting from a list of words or phrases), flow-chart completion, sentence completion and matching sentence endings.

The Academic Writing section contains two parts. Part one requires test takers to describe and explain a graph, table, chart or diagram presented in the prompts. Part two requires test takers to write an essay in response to a point of view, argument or problem.

Finally, the Speaking section consists of a 3-part interview between the test taker and an examiner. Test takers are required to express opinions and communicate information on everyday topics, experiences and situations, to speak at length on a given topic (without further prompts from the examiner), and then to express and justify opinions and to analyze, discuss and speculate about issues related to the topic of the long turn. Samples of all of the task types can be found on ielts.org.

The task types are described in more detail under the relevant section headings.

Typically, test-takers receive their test results 13 days after taking a paper-based test, five days after the computer test, and seven days after taking the IELTS Indicator.

Duolingo English Test (DET)

The Duolingo English Test (DET) is a computer-delivered, partially adaptive test that is described in its technical manual (Cardwell, LaFlair, & Settles, 2022 p. 3, hereinafter referred to as the DET Manual) as "a measure of English language proficiency for communication and use in English-medium." However, elsewhere it is described as a "high-stakes proficiency test that assesses English language proficiency for admission to English-medium universities" (Park, LaFlair, Attali, Runge, & Goodwin, 2022, p. 2). The test consists of both computer-adaptive (CAT) and non-CAT item types. In the five CAT types, performance on one item determines how difficult the next item will be; test takers typically encounter between four and six of each of these item types. There is a set number of non-CAT items in each administration, as described below. The five adaptive test item types are the following:

A. C-test (Read and Complete)[1]
The C-test task is based on the C-test developed by Raatz and Klein-Braley (1981), which is based on the notion that performance on a test with reduced redundancy (i.e., with some input missing) can provide evidence of general language proficiency (Klein-Braley, 1997). In a canonical C-test, the first and last sentences of an authentic passage are left intact, while the second half of every second word is deleted, starting with the second sentence (Klein-Braley, 1997), although McKay (2019) notes that this rule is often modified in practice. The sample items in the DET Official Guide for Test-takers[2] (Duolingo, 2021) and in practice tests mostly follow this canonical pattern, with some exceptions.

B. Visual yes-no questions (Read and Select)
In this item type, test takers are presented with a list of 18 words, some of which are actual English words and others not. The DET Manual calls this test "a variant of the 'yes/no' vocabulary test", which is intended to measure receptive vocabulary size.

The range of actual words in a given set is five to 13, based on the sample items in the DET Guide. Items are presented in groups of approximately equal difficulty.

C. Aural yes-no questions (Listen and Select)
The aural yes-no questions are similar to the visual yes-no questions, except that there are only nine words in each set, and the test taker has to click on each word to listen to it. The words are spoken by a mix of female and male voices with a Standard American accent. There is no limit to the number of times a test taker can click on an individual

1 The names in parentheses are used in communications directed at test-takers
2 Hereinafter, "DET Guide"

item, though each set is timed. For each set, between three and six of the nine words are actual words; distractors include non-words such as [fotogo] or [momər].

D. Dictation (Listen and Type)
This item type consists of single sentences such as “We have never spoken about work.” According to the DET Manual, dictation measures “test taker ability to recognize individual words and to hold them in memory long enough to accurately reproduce them” (p. 9). Test takers have one minute to complete the task, and they may listen to the sentence up to three times. The instruction appears on screen at the same time as the first iteration of the audio starts. Test takers must type the sentence into the box below the instruction.

E. Elicited Imitation (Read Aloud)
For this item type, test takers read a printed statement out loud. This task is not technically elicited imitation, as the term is generally understood in the literature (including the papers cited in the DET Manual), which involves listening to, processing, and then reconstructing a sentence, rather than simply using rote repetition (Jessop, Suzuki, & Tomita, 2007). This test task appears to target primarily intelligibility, rather than the ability to reconstruct a sentence using one’s internal grammar. The example sentences range from six to 15 words. Test takers have 20 seconds to record themselves with the computer’s microphone by clicking a button on the screen. They only have one chance to record their responses, and the test will advance to the next task when the time limit is reached.

The non-adaptive test types are the following:

F. Interactive Reading
The Interactive Reading sections of DET, described in detail in Park, LaFlair, Attali, Runge, and Goodwin (2022), involves five different tasks, all based on a single paragraph-length input text. Texts are either narratives or expository passsages, and each candidate encounters one of each type. Candidates have seven to eight minutes to complete all five tasks, which are the following:

1. Complete the Sentence: This is a modified cloze-type task, where several words in the passage are gapped and test takers much choose from five options. Only the first half of the passage is used for this section.

2. Complete the Passage: In this task, the entire passage is displayed except that one sentence has been deleted from the passage. Test takers choose the best sentence from four options presented.

3. Highlight the Answer: Test takers see the entire passage and a comprehension question. Rather than answering the question, test takers highlight the portion of the text that contains the answer.

4. Identify the Idea. Test takers are given four propositions and asked to identify the one that is contained in the passage.

5. Title the Passage. Test takers are given four alternative titles and asked to select the best one for the passage.

While the tasks themselves resemble those found in many other high-stakes reading tests, one innovation in this section is that all passages and test items are automatically generated, thus obviating the need to identify and edit existing texts to fit the specifications and facilitating the rapid generation of a large number of items. All items go through a human review process, described in Park et al. (2022). However, examination of these items in practice tests reveals some issues that could benefit from additional human review. For example, in one passage we encountered, the word following “A few” is gapped, and “few” is one of the options to fill the gap. Later in the passage, the test

taker encounters the phrase “kindness, courtesy and” with the following word gapped; again, “courtesy” is one of the options.

G. Extended Speaking and Writing

In addition to these CAT item types, which can vary in number from test to test, each test taker responds to four extended Speaking tasks: one picture description and three prompt-based tasks (two written and one aural). These items are calibrated for low, intermediate, and high proficiency, so that items of the appropriate difficulty can be delivered to test-takers based on their estimated ability from the CAT items. The Speaking items provide 20 seconds for preparation, and test takers are required to speak for at least 30 seconds, with a maximum of 90 seconds. Test takers also respond to an additional ungraded speaking prompt. This response is made available to institutions along with test scores.

In the extended Writing tasks, each test taker responds to three picture description tasks and one prompt-based task. The picture description tasks instruct test takers to write at least one sentence, and the prompt-based task requires at least 50 words. Test takers have five minutes to write and must produce at least 50 words before being able to submit the answer. In addition, test takers respond to an additional writing prompt, which instructs them to write for three to five minutes on the topic provided. Prior to 2022, this task was unscored and simply served as a writing sample to be provided to institutions; however, now it is scored and contributes to the overall score, while still being made available to institutions. Note that, for both the extended Speaking and Writing tasks, the number of tasks is fixed per test, unlike the computer-adaptive tasks.

All items are scored automatically. The items are not weighted in the traditional sense or combined to provide a total score; rather, the difficulty of each item is calibrated based on its lexical and syntactic characteristics, and performance on each item in the adaptive portion of the test contributes to an estimate of ability. When enough items have been administered to stabilize this estimate within given parameters, the final estimate is converted to the ability scale, which goes from 10 to 160 and is reported in intervals of 5 (see Settles et al., 2020 for further details).

Recently, Duolingo began reporting subscores on the same scale as the overall scale. Rather than reporting scores by skill (i.e., speaking, listening, reading, writing), Duolingo reports the following subscores: Conversation, Literacy, Comprehension, and Production, each subscore combining two of the traditional four skills according to the scheme in Figure 1 below. For example, any task that involves listening contributes to both the conversation and the comprehension score. According to LaFlair (2020), these subscales derive from a multi-dimensional scaling of more than 100,000 tests, along with a factor analysis of the same data, resulting in a general language factor and two dimensions, along which the test tasks can be placed (see Figure 1, adapted from LaFlair, 2020).

Figure 1. *DET subscores*

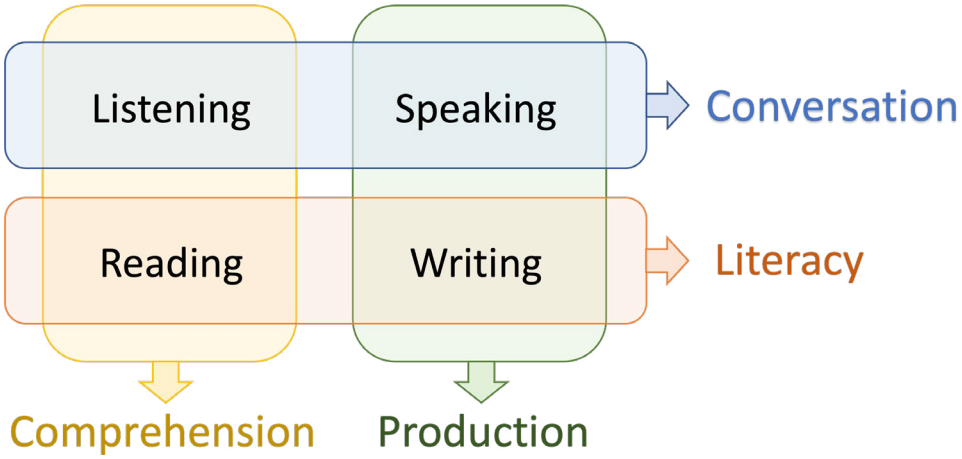


Table 1 below summarizes the contribution of each task type to these subscores.[3] Note that each task contributes to two subscores, depending on whether the modality (receptive vs. productive) or the channel (aural vs. visual) is invoked. As the table shows, more emphasis is placed on comprehension than production, and on the aural channel than the visual one, at least in terms of the task types.

Table 1. *Summary of task types and subscores*

	Format	Literacy (ability to read and write)	Comprehension (ability to read and listen)	Conversation (ability to listen and speak)	Production (ability to write and speak)
C-test	Adaptive	☑	☑		
Visual yes-no questions	Adaptive	☑	☑		
Aural yes-no questions	Adaptive		☑	☑	
Dictation	Adaptive		☑	☑	
Elicited imitation	Adaptive		☑	☑	
Interactive reading	Fixed number of tasks	☑	☑		
Extended Speaking	Fixed number of tasks			☑	☑
Extended Writing	Fixed number of tasks	☑			☑

We now turn to an examination of the validity of the two tests.

3 While Park et al. (2022) state on p. 5 that Interactive Reading contributes to these subscores, the same document later states that it will "ultimately be included" (p. 13) pending additional analyses.

III. Test taker characteristics

As noted earlier, an important consideration in the SCF is the test taker population: what their demographic and personal characteristics are, their purpose for taking a test, and how they intend to use the language. Data about population (test taker, score user) for both tests were collected from published materials and the tests' official websites.

For IELTS Academic, the test takers are those who want to "study in an English-speaking environment or university (higher education)"; one can also take IELTS Academic "for professional registration purposes" (see ielts.org). IELTS Academic scores have been used for admission purposes by more than 11,000 academic institutions mainly in Australia, Canada, New Zealand, and the UK.

According to the DET Manual, scores on the DET "are intended to be interpreted as reflecting test taker English language ability and used in a variety of settings, including for university admissions decisions" (p. 3). The DET has gained increasing popularity among test takers who wish to study and work in English-speaking countries, as it is less expensive and can be taken at home online. Many test takers have used it as an alternative to IELTS after the IELTS testing centers closed due to COVID-19.

Table 2 lists the general information about the test-takers and score users of the two tests, taken from the tests' respective websites.

Table 2. *IELTS and DET test taker profile*

	IELTS Academic	DET
Target levels (CEFR)	A1–C2	A1–C2
Purpose of test	To demonstrate readiness for academic study in the English medium	General proficiency, though marketed primarily as an admissions test for higher education
Test taker demographics	(2019 data*) Most frequent L1s: Arabic, Indonesian, Singhalese, Azeri, Chinese Total number of L1 not available Test taker age information not available	(2022 data**) 149 L1s Most frequent L1s: Mandarin, Spanish, English, Telugu, Arabic, Hindi, and Portuguese 81% of the DET takers are between 16 and 30 years of age Administered to test takers from 213 countries
Score use	The IELTS Academic test is suitable for those wanting to study in an English-speaking environment or university (higher education), as well as for professional registration purposes.	The test scores are intended to be interpreted as reflecting test taker English language ability and used in a variety of settings, including for university admissions decisions.
Number of institutions accepting score	IELTS is accepted by 9,736 educational institutions (including universities, colleges, and training programs) worldwide, according to its website. Detailed breakdown is not available.	1,936 undergrad schools 1,055 grad schools 483 secondary schools 424 other institutions

*Summarized from source data on the Test Statistics page on the IELTS official site: https://www.ielts.org/for-researchers/test-statistics
**Summarized from the DET Manual

Duolingo reports test performance by subscore and total score, including mean, standard deviation, and the 25th, 50th, and 75th percentile, respectively. There is no breakdown of scores by gender, first language, or other relevant demographic characteristics. IELTS, on the other hand, reports mean scores only, providing breakdowns by gender, nationality, and first language but not an overall mean score. There is no easy way to compare the performance on the two tests, except that the mean score on DET tests administered

over a recent one-year period is 108.79, four points higher than reported in LaFlair & Settles (2020) and the mean on IELTS is approximately 6.07[4]. According to Duolingo's concordance table (to be discussed in greater detail later in this report) this suggests that the two test taker populations are approximately equivalent in overall proficiency, with 50% of DET scorers scoring between 95 and 125, which represents bands 6 through 7.5 on IELTS. IELTS does not provide any information about the range of scores, however, so it is difficult to assess how similar the two populations are except for by comparing the mean scores.

IV. Cognitive and context validity

In this section of the report, we look at the four main language skills (reading, listening, speaking, and writing) as they are assessed in the two tests. We are combining cognitive and context validity in this section, as the contextual variables are often determinative of the kinds of cognitive processes elicited by test tasks. As noted earlier, IELTS has sections dedicated to each skill, while DET tasks are intended to measure integrated skills, as defined by Duolingo. For each section, we outline the cognitive processes involved in successful language use, and the most relevant contextual variables that mediate these processes. We then provide a comparison of these variables between the two tests.

In our analysis, we have tried to follow closely the process set forth by Taylor and Chan (2015) in comparing five different tests. For each skill area, they developed a pro-forma template, which was filled out for each test. Our tables below are adapted from their pro-formas (found in appendices 5 through 8 of their report), and in the case of IELTS, we have relied on much of their analysis.

Reading

As noted earlier, we are focusing our comparison on the IELTS Academic Reading rather than the General Reading. The main difference between the Academic and General versions of IELTS is the nature of the texts, though not the question types, encountered by test takers.

Following Taylor and Chan (2015), we provide a test task analysis based on the SCF and processing models outlined in Weir (2005) and Khalifa and Weir (2009) to determine the degree to which the cognitive processes elicited by test tasks involving reading are similar to those used in reading in non-test situations.

The IELTS Reading section consists of 40 comprehension items based on three passages, with 60 minutes to complete the section. For the purposes of this analysis, we are only considering three Duolingo tasks: C-test (Read and Complete), visual yes-no questions (Read and Select) and the new Interactive Reading; see the description above for these task types. Table 3 summarizes the tasks and the reading skills assessed by both tests. It is evident from the table that IELTS requires much more extended reading than DET and targets a wider range of reading skills and purposes. While the sample IELTS Reading passages we reviewed were all at least 250 words long, and some were more than 800 words, the longest reading passages are those in the Interactive Reading sections. Of the four passages we examined, the longest was 126 words. We go further into these issues by referencing Khalifa and Weir's framework.

4 The mean for females is 6.104 and for males is 6.039; as noted, no overall mean score is reported.

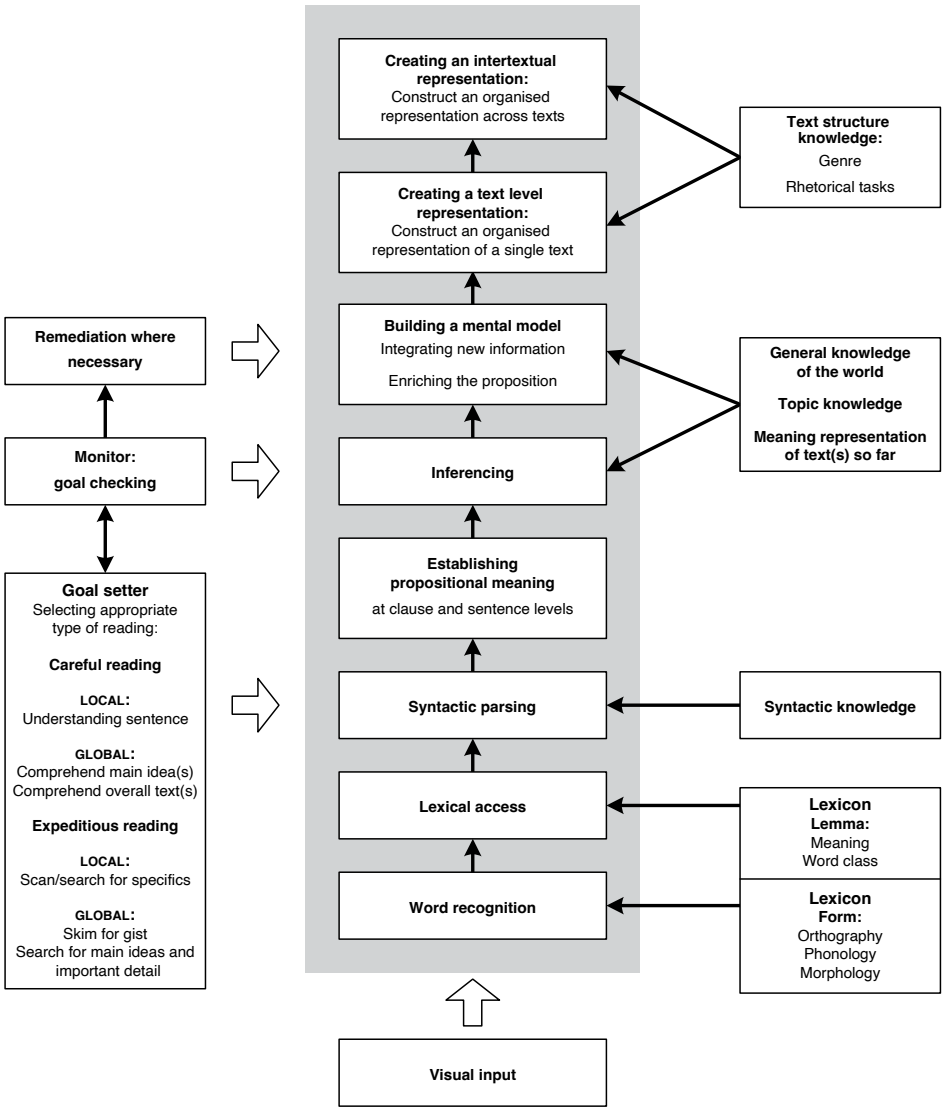
Table 3. *Comparison of IELTS and DET Reading tasks*

	IELTS	DET
Task description	40 questions based on three passages, averaging approximately 800 words in length Mainly selected response (multiple choice, true/false/not given, some short answers, e.g., fill in the blanks)	**Read and select:** Candidates determine whether a string of letters is a real English word or not. Words are presented in blocks of 18. **Read and complete:** Candidates read and complete a short (approximately 30–50 word) modified C-test. **Interactive Reading:** Candidates respond to six item types based on a paragraph-length text.
Timing	60 minutes to read three passages and answer 40 questions	**Read and select:** 1 minute per 18-word block **Read and complete:** 3 minutes per passage **Interactive Reading:** 7 or 8 minutes per passage
Skills focus	Items are targeted at the following skills: Read for the general sense of a passage Read for main ideas and details Understand inferences and implied meanings Recognize writer's opinions, attitudes, and purposes Follow the development of an argument	**Read and select:** Distinguish English words from non-words **Read and complete:** Use lexical and syntactic knowledge to complete gapped words in a short passage **Interactive Reading:** Five item types as follows: Complete sentences with gapped words Complete paragraphs with gapped sentence Locate the answer to a comprehension question Choose the idea that is present in the text Choose the best title for the text
Weighting	All items weighted equally.	Unclear; all items contribute to estimate of candidate ability

Cognitive and context validity

In investigating the cognitive and context validity of the two tests, we first lay out the cognitive processes involved in real-world reading, and then discuss the contextual variables that mediate successful cognitive processes. Figure 2, from Khalifa and Weir (2009), outlines the types of readings, cognitive processes, and knowledge sources involved in successful reading. In presenting this model, Khalifa and Weir's goal is to identify generic reading processes that potential test candidates would engage in during real-world reading, so that these processes can be sampled in a reading test. These factors form the basis for comparing reading tasks on the two tests.

Figure 2. *Types of reading and their associated cognitive processes (Khalifa & Weir, 2009, p. 43)*



The leftmost column in the figure includes the metacognitive activities of goal setting, monitoring, and remediation. For our purposes, the key factor is the selection of an appropriate type of reading. Reading scholars have conceptualized reading for different purposes along two axes: expeditious vs. careful reading, and local vs. global reading. Expeditious reading involves reading a text as efficiently as possible to either get a gist of the entire text (skimming—global), to locate information on a particular topic (search reading), or to locate a specific piece of information such as a word or number (scanning—local). Careful reading, on the other hand, involves attempting to extract meaning to comprehend an entire text (global) or to establish the meaning of a single proposition within a text (local).

While most reading tests focus on careful reading for comprehension, Khalifa and Weir note that expeditious reading (reading quickly and selectively to find specific information or get a sense of the overall gist of a passage) is often more problematic for students than careful reading, and thus should not be excluded from the test construct.

The cognitive processes involved in reading are found in the center of the diagram. Local reading involves the processes of word recognition, lexical access, syntactic

parsing, and establishing the literal meaning of a clause or complete sentence. Beyond these basic processes, global reading (reading for comprehension beyond the sentence) involves the processes of inferencing, building an ongoing mental model of the text so far, and then creating a model of the text as a whole. Finally, when more than one text is involved, readers create intertextual representations, synthesizing and reorganizing information received from multiple texts. The degree to which these processes are invoked in reading for a test is referred to as **cognitive validity.**

The right-hand column of the text refers to the kinds of knowledge brought to bear in reading comprehension that mediate the cognitive processes in reading. At the lowest levels, knowledge of lexis and syntax are the primary knowledge sources involved. As texts and reading tasks become more complex, readers engage their knowledge of the world, knowledge of the topic, and what they have understood from the text so far to make inferences and build a mental model of propositions in the text. To create a representation of the text as a whole, or to synthesize multiple texts, readers rely on their knowledge of how texts are ordinarily structured, both in terms of genre (e.g., email, textbook, blog post) and rhetorical task (e.g., narration, persuasion).

The parameters of a test that mediate cognitive processes and impact test performance fall under the rubric of **context validity**. Aspects of context validity that are important for reading include features of the task setting and the linguistic demands of task input (what the candidate reads) and output (the expected response to the input). Khalifa and Weir mention the following under the rubric of task setting:

- Response method (selected vs constructed responses of various sorts)
- Weighting (how different tasks or items factor into the final score, which may influence candidates' goal-setting processes)
- Knowledge of criteria (whether candidates know how items will be scored, which is perhaps less relevant for reading than for productive skills, particularly in selected response items)
- Order of items (particularly with respect to whether the order of items parallels the order of information in a text)
- Channel of presentation (this refers to whether the text includes visuals or other non-verbal information)
- Text length (longer texts allow for a wider range of reading skills, such as distinguishing main ideas from details)
- Time constraints (the speededness of the test).

Factors related to the linguistic demands of the input include the following:

- Overall text purpose (to inform, persuade, convey emotion, entertain, keep in touch)
- Writer-reader relationship
- Discourse mode (including genre and rhetorical task, e.g., inform or persuade)
- Functional resources
- Grammatical resources (complexity at the sentence level and the phrase level)
- Lexical resources (word frequency, other measures of vocabulary sophistication)
- Nature of information (concrete vs. abstract)
- Content knowledge (relationship between text content and reader's background knowledge).

We now provide a more in-depth comparison of cognitive and context validity considerations for IELTS and DET.

Cognitive validity in IELTS vs. DET

Table 4 provides a summary of the cognitive processes elicited by the reading tasks in the two tests. As the table shows, there are major differences between the two tests in terms of cognitive processes. IELTS test tasks elicit both expeditious and careful reading, both locally and globally, while DET requires careful reading at the local level only. To answer IELTS Reading items, candidates must establish propositional meaning, make inferences, build a mental model of the text, and create a text-level representation of passages up to several paragraphs long. The DET Interactive Reading targets similar skills, but since the texts consist of a single paragraph, the items only minimally evoke global and expeditious reading. Similarly, while test takers may need to build a mental model of a text to select the best title, this model will not be as complex as that required by IELTS passages. In contrast, correct answers to DET C-test items require only word recognition, lexical access, and syntactic parsing.

Table 4. *Cognitive validity in reading: IELTS vs DET*

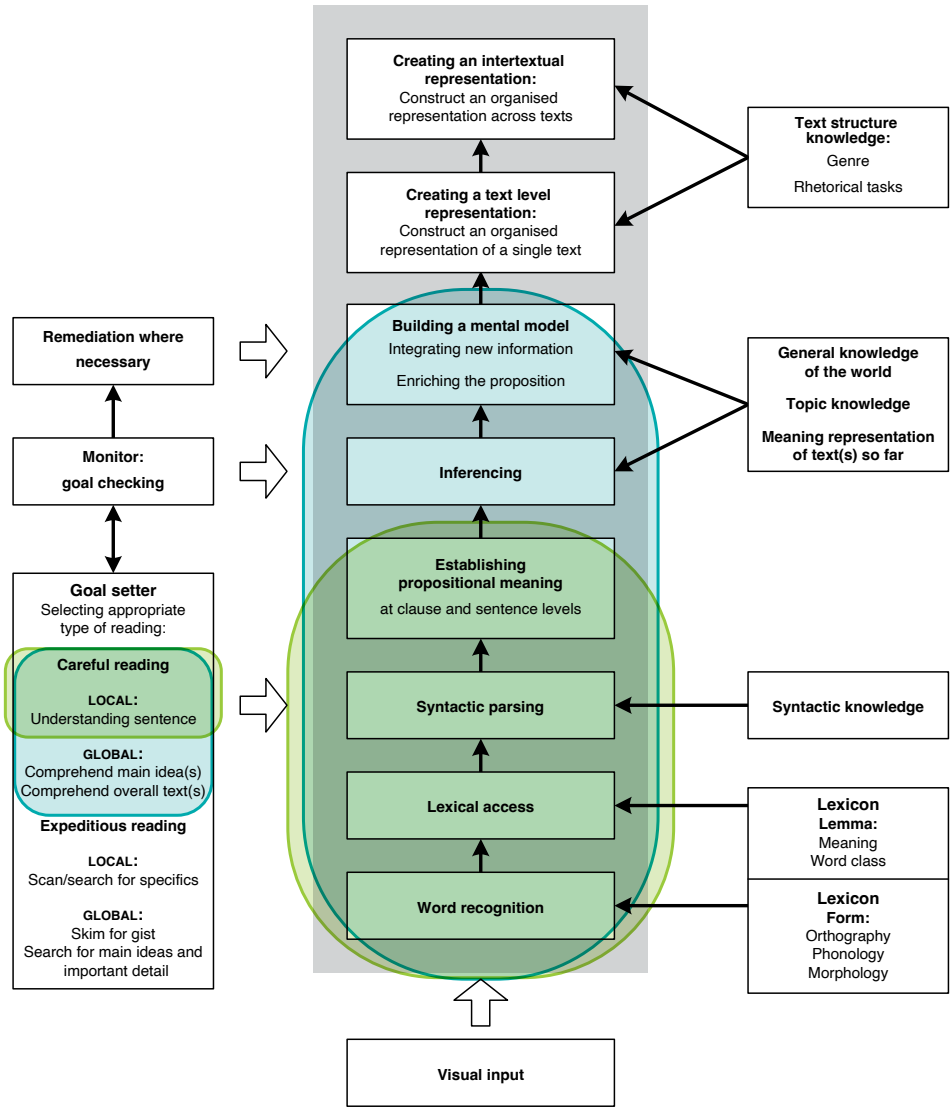
	IELTS	DET
Goal setting	**Expeditious reading: Local***	Expeditious reading: Local
	Expeditious reading: Global	Expeditious reading: Global
	Careful reading: Local	**Careful reading: Local**
	Careful reading: Global	Careful reading: Global
Levels of reading	**(Word recognition)****	**Word recognition**
	(Lexical access)	**Lexical access**
	(Syntactic parsing)	**Syntactic parsing**
	Establishing propositional meaning	**Establishing propositional meaning**
	Inferencing	**Inferencing**
	Building a mental model	**Building a mental model**
	Creating a text-level representation	Creating a text-level representation
	Creating an intertextual representation (multi-text)	Creating an intertextual representation (multi-text)

*Bold face indicates the skills that are invoked in responding to test items.

**Skills in parenthesis are not directly tested, but are enabling skills to complete other reading tasks.

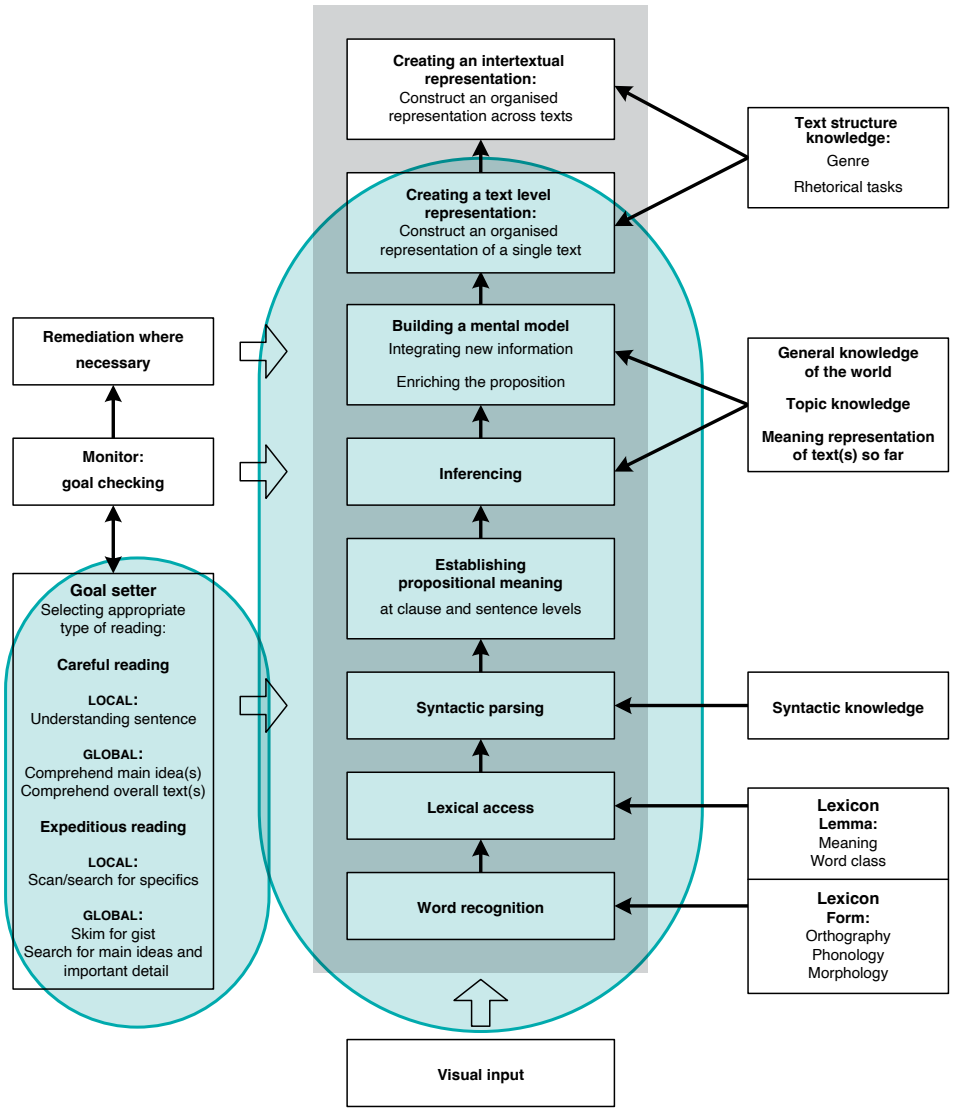
In terms of construct coverage, therefore, we can see that IELTS covers the reading construct much more fully than DET, as illustrated in Figures 3 and 4 below.

Figure 3. *DET construct coverage*



NOTE: The green area represents construct coverage prior to the addition of Integrated Reading; the blue area represents additional construct coverage (albeit in very short passages) with the introduction of Integrated Reading.

Figure 4. *IELTS construct coverage*



Context validity

As stated earlier, the contextual variables in a reading test mediate the cognitive variables discussed in the previous section; that is, the degree to which cognitive processing is made easier or more difficult is determined in large part by aspects of the test task. Table 5 below presents a comparison of relevant contextual variables as presented earlier. We have relied in part on Taylor and Chan's (2015) analysis for some of the factors in IELTS, confirming them with our own independent evaluation of the sample texts and items provided by IELTS. For the DET, we independently examined the C-test sample passages found in the DET Guide, along with the passages encountered in the two practice tests described earlier, and came to a consensus where there was any disagreement. The visual yes-no questions are not included in this analysis, as they consist of single words only. We also examined four Interactive Reading passages: one from the DET Manual, two from the DET Guide, and one from a practice test. However, these passages were not labeled by level so it is not clear what level of test taker they target.

Table 5. *Analysis of IELTS and DET context validity variables for reading*

	IELTS	DET
Domain	Social Work **Academic**	**Social** Work **Academic**
Discourse mode	Descriptive **Historical/biographical** **Expository** **Argumentative** Instructive	**Descriptive** **Historical/biographical** **Expository** Argumentative Instructive
Content knowledge	General More general than specific **Balanced** More specific than general Specific	General **More general than specific** Balanced More specific than general Specific
Nature of information	Only concrete **Mostly concrete** Fairly abstract Mainly abstract	**Only concrete** Mostly concrete Fairly abstract Mainly abstract

As the comparison shows, the nature of the texts is somewhat more academic in IELTS, particularly in terms of the discourse mode, use of non-verbal information, and level of abstractness in the texts. For example, DET includes descriptive passages, whereas IELTS includes argumentative passages. IELTS also requires test takers to combine verbal and visual information, while DET does not.

Turning now to a discussion of linguistic variables in the two tests, we provide an analysis of the lexical and syntactic features of the input texts. Again, we are relying for this analysis on the sample IELTS passages and, for the DET, on the C-test and Interactive Reading passages described above.

Syntactic complexity

Using Lu's (2010) Syntactic Complexity analyzer, we calculated several variables related to syntactic complexity, as follows. For the DET C-tests, the three passages at each level were combined into a single text for analysis. In addition, the four C-tests that appeared in two official practice tests described above were combined into a single text, to represent the entirety of reading that is involved in a practice test. Nine IELTS Reading passages from among those provided by IELTS in their official test preparation materials were also analyzed. Unlike DET C-test passages, to the best of our knowledge, IELTS passages are not explicitly targeted at specific proficiency levels; rather, test takers at all levels encounter the same passages, some of which may be easier to read than others. To account for the wider range of texts in IELTS, we report the median, minimum, and maximum for each measure.

Selected variables related to syntactic complexity are found in Table 6. As the table shows, the IELTS passages are overall more complex, with longer sentences and clauses, and more coordinated phrases and complex nominals per T-unit, defined as a main clause with any subordinate clauses that may be attached to it (Hunt, 1964). For more detailed discussion of measures of syntactic complexity, see Lu (2010).

Table 6. *Comparison of syntactic variables across readings in the two tests*

	Total number of words*	Mean length of sentence	Mean length of clause	Clauses per sentence	Complex T-unit ratio**	Coordinate phrase per clause	Complex nominal per clause
DET Level 1 (D1) C-test	161	11.50	8.47	1.35	0.27	.16	0.90
DET Level 2 (D2) C-test	117	11.70	10.64	1.10	0.10	.36	0.91
DET Level 3 (D3) C-test	157	17.44	11.21	1.56	0.44	.14	1.79
DET Practice test 1 (DP1) C-test	236	13.89	11.90	1.18	0.11	.35	1.65
DET Practice test 2 (DP2) C-test	224	14.00	10.67	1.31	0.31	.28	1.33
DET IR (Median)	110	16.45	9.17	1.80	0.55	0.35	1.28
IELTS Low (IL) C-test	218	21.00	10.31	1.38	0.31	.16	1.39
IELTS Median (IM) C-test	434	22.79	12.00	1.95	0.50	.35	1.71
IELTS High (IH) C-test	887	30.25	17.39	2.85	0.70	.47	2.51

*For DET, the C-test passages are combined for this analysis. The average passage length is: D1 (53.67), D2 (39), D3 (52.33), DP1 (59), DP2 (56).

**Complex T-unit ratio is defined as the ratio of complex T-units (T-units containing more than one clause) to total T-units, and is a measure of subordination.

As the table shows, for many of these syntactic complexity indices, the Level 3 DET C-test passages fall between the lowest and the median for IELTS passages. The mean length of sentence is shorter for all DET C-test passages than for the IELTS passages. The results for the DET Interactive Reading are similar. This finding suggests that IELTS candidates, on average, will encounter and need to parse and comprehend sentences of higher complexity than DET candidates.

Lexical complexity

There are numerous ways to describe the words in a given text: some of the most frequent measures have to do with word frequency, lexical density, and other measures of word sophistication. We look here at these factors in the texts encountered by IELTS and DET test takers. For the lexical analysis, we put the same texts as described above into the VocabProfiler (https://www.lextutor.ca/vp/eng/) to check the overall lexical level of the texts. This tool analyzes the words in a text according to whether they appear in the most frequent 1,000 word families of English (K1), the second 1,000 (K2), the Academic Word List (AWL) compiled by Coxhead (2000), or not on any of these three lists (off-list).

Furthermore, because the DET C-test involves completing targeted gapped words, we wanted to analyze the vocabulary level of the gapped words. We therefore did a further investigation of these words in the sample passages, and those presented in the two practice tests. Using the output from VocabProfiler, we entered each gapped word beyond the K1 level into the American version of the English Profile (https://www.englishprofile.org/american-english) to find its CEFR level (i.e., where learners begin to use it productively). Some words were not found in the English Profile; in all cases, however, these words are repeated elsewhere in the passage, so test takers do not need to recall them from memory. As Table 7 shows, the vocabulary gets consistently more sophisticated from Level 1 to Level 3; in the two practice tests, the number of

off-list words was substantially higher than in the sample items, and the percentage of academic words lower, but the passages were presumably of mixed difficulty. However, approximately half of all gapped words are function words, and very few gapped words are not in the 1,000 most frequent word families, and very unfamiliar or text-specific words (e.g., dengue, opera) tend not to be gapped unless they appear elsewhere in the text. Thus, the C-test appears to target function words more than content words, making it perhaps more of a test of syntactic knowledge than lexical knowledge.

Table 7. *Vocabulary range of DET C-test passages (distribution of word frequencies for total passages and CEFR levels of gapped words beyond the 1K level)*

	Level 1	Level 2	Level 3	Practice test 1	Practice test 2
K1 words	87.5%	79.49%	69.48%	74.04%	74.63%
K2 words	2.5%	5.98%	10.39%	3.83%	4.88%
AWL words	.62%	5.98%	9.74%	1.70%	3.41%
Off-list words	9.38%	8.55%	10.39%	20.43%	17.01%
Content words gapped	43%	56%	48%	38%	46%
K2 words gapped	freezing (B1)	birth (A2)	combined (B2) electrical* (B1) satisfied (B1)		modesty (C1) treatment (B2)
AWL words gapped	environment (B2)	migrate (N/A) psychological (B2)	community (B2) vision (C1)	predominantly (C2)	conducts (C2)
Off-list words gapped		symptom (B2) medications (C2)	opera* (A2)	orbits* (N/A) vulture* (N/A) oval (B2)	dengue* (N/A)

*These words are found elsewhere in the passage.

Table 8 shows the vocabulary range of the vocabulary in the sample IELTS Reading passages and the DET Interactive Reading passages. Comparing the median IELTS text with the DET sample items, the IELTS texts appear to fall in the range between Level 2 and Level 3 of the DET items, suggesting that the vocabulary presented to candidates may be slightly more advanced at the highest level of the DET. However, as noted above, DET C-tests tend not to focus on advanced vocabulary in the items that are gapped.

Table 8. *Vocabulary range of IELTS Reading and DET Interactive Reading passages**[5]

	IELTS			DET		
	Minimum	Median	Maximum	Minimum	Median	Maximum
K1 words	64.35%	72.15%	85.14%	68.69%	77.15%	84.94%
K2 words	3.91%	6.21%	8.75%	1.01%	2.05%	10.68
AWL words	3.85%	6.57%	13.06%	0%	6.58%	18.18
Offlist words	3.15%	14.45%	21.45%	2.91%	12.41%	17.24%

*Based on 9 IELTS passages and 4 DET passages

In summary, a detailed comparison between the IELTS Academic Reading and the Reading section on the DET reveals that the DET assesses a much narrower range of reading purposes and cognitive processes, with only the C-test and Interactive Reading tasks coming close to assessing reading as it is generally defined by scholars. The new

5 IELTS is based on nine sample passages; DET IR is based on four.

Interactive Reading items expands the reading construct assessed in DET and is more clearly academic in nature than the C-test items. However, we had access to only four texts and their associated items, so it is difficult to make generalizations about the lexical and syntactic features of the Interactive Reading texts. The C-test tasks at the highest level tend to be syntactically somewhat less complex than IELTS Reading passages, and of course much shorter. The vocabulary in the highest level tends to be somewhat less frequent or familiar than that in typical IELTS passages; however, the C-test tends not to assess these less frequent words (i.e., they are not gapped). Finally, the texts presented in IELTS reflect academic texts more closely, in the sense that they include argumentation, a very common rhetorical mode in academic writing, and are often more abstract.

Listening

As with the reading analysis, we follow Taylor and Chan's (2015) framework for the listening analysis, which includes task features (including cognitive criteria), features of input, and language content. This model was expanded and modified from the evaluating criteria for the reading task to take into consideration the specific skills involved in processing audio input (Field, 2013).

Unlike the Reading and Writing tasks, IELTS Listening does not distinguish between academic and general training purposes. The four parts in the IELTS Listening section cover both social and academic contexts in the format of conversations and monologues.

The IELTS Listening section consists of 40 comprehension questions based on four passages, on average one question per 75 words. The question types include the following: three-option multiple choices, gap/form completion, labeling diagrams and maps, multiple matching, and choosing from a list. Responses in gap-filling are a maximum of three words. All question types may appear with any of the passages. The test-takers are instructed to read the questions first before the audio passages start. The audio passages are of various lengths with an average of 152 seconds (based on the sample IELTS passages published on the official website), and the test takers have 30 minutes to complete all questions.

Since the DET tasks do not separate a listening section on its own, we reviewed the relevant tasks that include a listening element: aural yes-no questions (Listen and Select) and dictation (Listen and Type). Although there is an aural prompt for some of the extended Speaking tasks, listening is not the main construct and is not graded; therefore, we did not include the DET extended Speaking task in our analysis of listening items. Table 9 summarizes the tasks and the skills assessed in these tasks.

Table 9. *Comparison between IELTS and DET Listening*

	IELTS	DET
Task description	Four passages with 10 questions each. A variety of question types is used. Listening passages are approximately 269 words in length on average.	**Aural yes-no:** Candidates determine whether a string of sounds is a real English word or not. Words are presented in blocks of nine. **Elicited imitation:** Test takers listen to a spoken sentence and transcribe it.
Timing	Passages are 152 seconds long on average. Around 30 minutes for the entire Listening section.	**Aural yes-no:** 90 seconds to identify real words among nine spoken words/ pseudowords. **Elicited imitation:** One minute per item.
Presentation	Audio is played one time only. Candidates can preview the items for 45 seconds before audio is played.	**Aural yes-no:** no limit on repeating the audio within the time limit. **Elicited imitation:** Audio can be repeated up to three times.
Skills focus	Understand main ideas and detailed factual information. Understand the opinions and attitudes of speakers. Understand the purpose of an utterance. Follow the development of ideas.	**Aural yes-no:** Distinguish words from non-words presented aurally. **Elicited imitation:** Transcribe sentences presented aurally; hold input in short-term memory long enough to transcribe sentence.

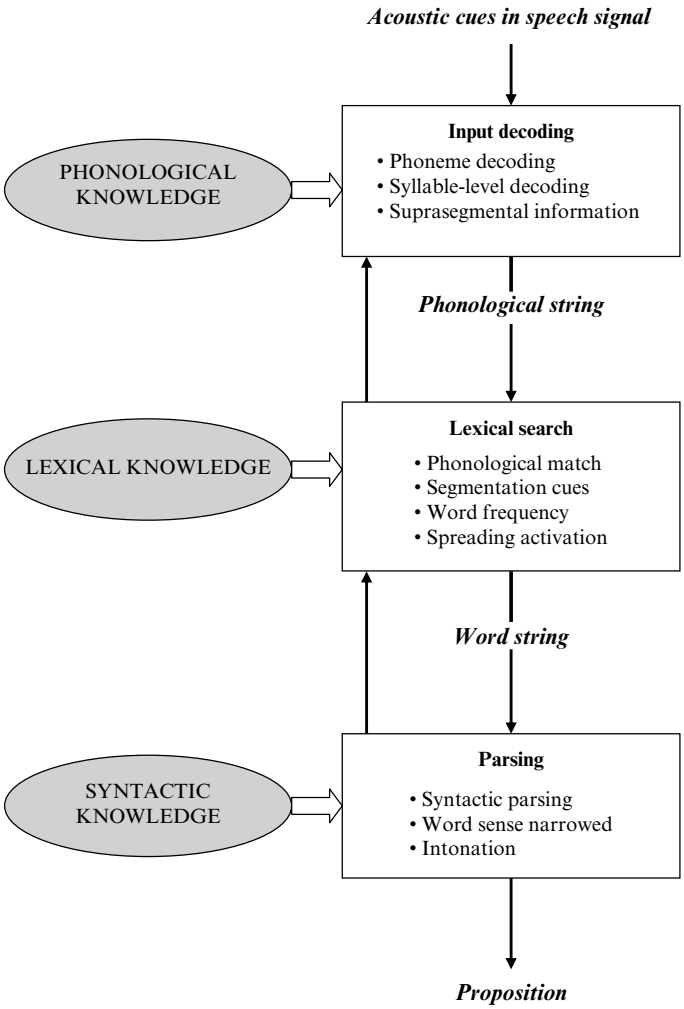
As seen in Table 9, the IELTS Listening section assesses a wider range of listening skills with more extended listening input for test takers to process. IELTS Listening tasks also require a higher level of retention of information from listening to each passage only once, while the DET Listening tasks typically allow test takers to listen to the materials more than once. The difference in task types and task features addresses different levels of underlying cognitive processes, which we elaborate on in the following sections.

Cognitive processing

Field's (2013) cognitive model of listening includes five levels of processing: input decoding, lexical search, parsing, meaning construction, and discourse construction. These levels illustrated the cognitive progress from recognizing individual sounds to pragmatic decoding of utterances in the discourse and social context. The framework we adopted to compare the two tests takes into consideration how task features as well as external factors may influence these cognitive processes. Although the core concept of cognitive validity does not differ from that of other skills such as reading, writing, and speaking, Field highlights the fact that listening is far more complex than the other skills.

Specifically, according to Field (2013), the first three levels – input decoding, lexical search, and parsing – are referred to as lower-level processes, which are invoked when listeners encode input signals into language. Figure 5 illustrates the principal processes and the linguistic knowledge source used to support the three levels. Although the graphic representation of these levels seems sequential, it is worth noting that recent research has shown that one or more levels in these processes can occur simultaneously. Moreover, the upwards arrows linking the three processes indicate an overruling effect for positive information from a higher level to cancel out negative information at a lower level.

Figure 5. *Model of lower-level processes in listening, from Field (2013), drawing upon Cutler and Clifton (1999) and Field (2009)*

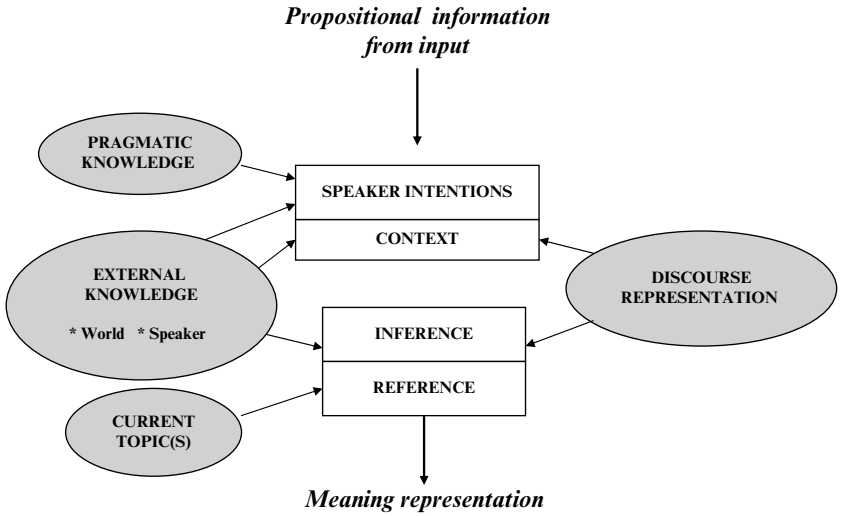


For the steps listed at the input decoding level, it has been argued that the core of this process is the syllable-level decoding (Field, 2013, p. 96). By incorporating phonological knowledge while decoding the cues in the input, the listener arrives at an output that is beyond a string of phonemes, but constructed with suprasegmentally marked syllables. The lexical search process requires the listener to draw upon lexical knowledge to identify a string of words – or a range of senses to be confirmed at a later stage – from the sequences of sounds from the input decoding level. All information is processed online at the parsing level while the utterance is being produced. As shown in Figure 5, this level includes not only syntactic parsing, where the listener imposes a syntactic structure to the words, but also narrowing down the previously identified range of word senses to suit the appropriate context. Meanwhile, at this stage, the intonation contour of the utterance becomes available for constructing a literal proposition, foregrounding the meaning and discourse building during higher-level processing.

Higher-level processes start with the literal, abstract, and context-independent proposition produced by the lower-level processes. These processes can be best understood with consideration of what typically happens in speech processing. In a conversation, listening is made easier by the fact that most utterances are short, with relatively simple syntactic structure. At the same time, much of ordinary speech is encoded in highly abbreviated forms (deictics, contractions, etc.), leaving a more challenging task for listeners to connect the speech with referents available from the context and to infer meaning. As Field (2013) pointed out, “the raw meaning of the speaker’s words is insufficient to convey the significance of what is being said or why it has been said” (p. 100). The process of constructing propositional information derived from the aforementioned

lower-level processes is referred to as meaning construction. Figure 6 illustrates the model for meaning construction and the knowledge sources used to support these processes.

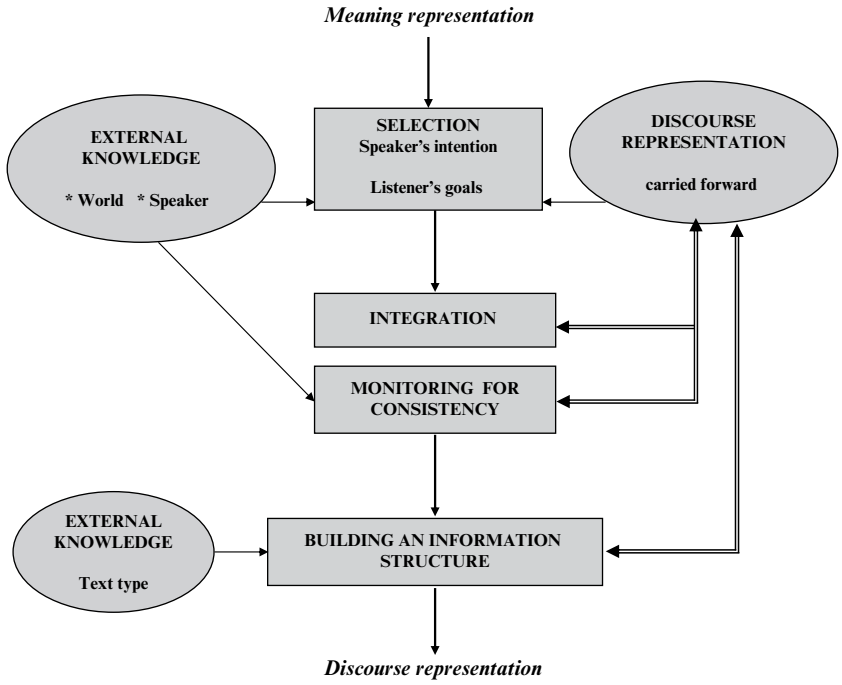
Figure 6. *Model of meaning construction in listening (Field, 2013)*



The boxes in the middle of the graph represent the types of information the listener needs to supply to enrich the literal propositional information constructed at the lower-level processes. The oval shapes to the left are knowledge required to accomplish these processes. Insufficiency of any supporting knowledge of skills to process the knowledge may lead to incomplete understanding of the message.

Beyond meaning construction, discourse construction as another high-level process requires the listener to take account of all relevant information and events to interpret the meaning representation from the meaning construction level. These processes are illustrated in the middle squares of Figure 7.

Figure 7. *Model of discourse construction in listening (Field, 2013)*



For example, the listener may select certain information that is relevant to the development of the current discourse for future use, detect implicit logical connections between pieces of information, evaluate previous judgement of ongoing discourse based

on new incoming information, and create a hierarchical structure of the main points being said in the discourse. As seen in Figure 7, interpretation of the meaning representation requires a similar range of knowledge types as required for meaning construction (Figure 6).

Cognitive validity

The final comparison using the above model is shown in Table 10, which shows that the two tests differ in their cognitive processes. The most obvious difference between the two tests is that the DET does not go beyond the level of parsing, while IELTS Listening items require meaning construction. Based on our analysis of the published sample tests, IELTS Listening questions typically target detailed information that has been framed with signaling language in the listening materials. Upon reading the questions prior to listening to the audio, test takers need to build an incomplete mental model and actively seek the missing information while decoding, interpreting, and selecting all the incoming pieces of information from test audios. In contrast, the DET Listen and Select items have a specific focus on input decoding and lexical search, for which test takers decode the input to match the phoneme, syllable, and suprasegmental features that are possible in English and identify whether they are existing words in the English language. The Listen and Type task in DET also targets similar skills, while potentially involving parsing, which facilitates the holding of words and phrases in short-term memory long enough to transcribe them. The DET tasks do not assess higher-level processing skills.

Table 10. *Cognitive validity of IELTS and DET Listening*

	IELTS	DET
Cognitive processing: Goal setting	Listen for specific information Constructing meaning in context Constructing a discourse model	Distinguish words from non-words Hold input in short-term memory long enough to transcribe sentence
Cognitive processing: Levels of listening targeted by items	**Input decoding*** **Lexical search** **Parsing** **Meaning construction** Discourse construction	**Input decoding** **Lexical search** **Parsing** Meaning construction Discourse construction

*Bold face indicates the skills that are invoked in responding to test items.

Context validity

We provide a detailed discussion of context validity in the Reading section, which is to a large degree applicable to the analysis of the Listening tasks, except that for the listening assessment, we also need to consider speaker characteristics associated with the audio input. Features that consider the speakers for context validity include:

- Speech rate
- Length of individual utterances (words between pauses) and entire listening text
- Variety of accent
- Sociolinguistic considerations
- Number of speakers.

Therefore, we organize our comparison into two categories: “input as text”, which considers the textual features of the input texts, and “input as recorded material”, which considers mostly the speaker features. Table 11 compares the key elements of input as text for context validity of the Listening tasks in the two tests. For the relevance of this

comparison, we select the DET Listen and Type question type, since the individual words in the Listen and Select task do not provide information relevant to this comparison.

Table 11. *Analysis of IELTS and DET context validity variables for Listening (input as text)*

	IELTS	DET
Domain	**Social** Work **Academic**	**Social** Work **Academic**
Discourse mode	**Descriptive** Historical/biographical Expository **Argumentative** Instructive	**Descriptive** Historical/biographical Expository Argumentative Instructive
Content knowledge	General More general than specific **Balanced** More specific than general Specific	General **More general than specific** Balanced More specific than general Specific
Cultural specificity	Neutral Mostly neutral **Balanced** Mostly culturally specific Culturally specific	**Neutral** Mostly neutral Balanced Mostly culturally specific Culturally specific
Nature of information	Only concrete **Mostly concrete** Fairly abstract Mainly abstract	**Only concrete** Mostly concrete Fairly abstract Mainly abstract
Presentation	**Verbal** Non-verbal (e.g., graphs, pictures) Both	**Verbal** Non-verbal (e.g., graphs, pictures) Both

In terms of input as texts, even though texts in both tests cover social and academic domains, IELTS Listening input is significantly longer and is presented in much more varied discourse modes. The amount of information test takers need to process to answer IELTS Listening questions is significantly denser and more context-specific than that required by the DET. The two scored listening task types in the DET, namely word recognition and dictation, do not require processing of longer texts, although, according to the DET Manual, the phonological features measured by these tasks are strong predictors of listening comprehension ability (p. 9).

Table 12 provides a summary of context validity for listening input as recorded materials (audios). For this analysis, again, only the DET elicited imitation question type was used. The DET Official Guide provided prompt sentences in the elicited imitation items at each level, which were analyzed and presented separately in the table.

Table 12. *Analysis of context validity variables for listening on both tests (input as recorded material)*

	IELTS	DET
Transmission	Audio. No visual support. Context and topic provided in instructions.	Audio. No visual support. No context or topic provided.
Articulation	Natural in intonation and pausing	Natural in intonation and pausing
Speech rate	Dialogue: 3.34 syllable/sec Monologue: 3.78 syllable/sec	Average for dictation task: Level 1: 4.55 syllable/sec Level 2: 4.78 syllable/sec Level 3: 5.48 syllable/sec
Utterance* and text length	Up to 16 words per utterance; average 270 words per passage	Levels 1 & 2: average 6 words per sentence Level 3: average 14 words per sentence
Accent	Range of accents: British, American, Australian/NZ	American
Gender	Mix of male and female voices	Mix of male and female voices
Lexical level**	K1–K3	Level 1: A1, some A2 Level 2: Mostly B2, some B1 Level 3: Mostly C2, some B2/C1
Grammatical level	A high level of syntactic complexity. A high proportion of subordination, both within utterances and across utterances.	Level 3 sentences include subordination, passive voice, and less common tense/aspect combinations.

*An utterance is defined as the words spoken between pauses in oral discourse. This measure is more relevant to IELTS than to DET, since the DET texts are individual sentences, not stretches of discourse.

**Lexical and grammatical levels for IELTS are those reported by Taylor and Chan (2015). Because DET items are discrete sentences rather than extended texts, we felt it was more appropriate to evaluate the lexical levels in terms of their CEFR level based on the English Profile.

In terms of audio input features, the IELTS test passages include a variety of speaker accents while the DET listened texts appear to include standard American accents only. Meanwhile, although the IELTS Listening texts are significantly longer, it appears that the DET Listening texts are spoken at a faster rate (for all three levels). The speech rates for the sentences (ranging from six to 14 words) in the DET dictation task are spoken at about five syllables per second, while both monologues and conversations in the IELTS test are about three to four syllables per second, at the length of 270 words per text.

Based on our analysis, we found a clear progression on the DET from Level 1 to Level 3 in terms of the vocabulary levels. As for grammatical level, there is a similar progression, with Level 1 and 2 texts consisting primarily of simple sentences and questions, and Level 3 texts containing less frequently encountered tense/aspect combinations such as future perfect (e.g., “They will have tried to talk to you by the time the story has published [*sic*].”).

To summarize, the listening construct for DET appears to be based solely on intensive listening for vocabulary and grammatical structure, rather than listening for meaning or to understand the gist of extended discourse, which are essential skills for academic success. Thus the listening construct appears to be the most under-represented of all the skills in DET, compared with IELTS.

Speaking

The Taylor and Chan (2015) framework for evaluating the cognitive and context validity of speaking tests, which we follow here, was adapted from Weir (2005) and Field (2011). The IELTS Speaking test consists of an 11–14-minute face-to-face oral interview with a trained examiner. The DET requires speaking in three task types: reading aloud, extended Speaking, and the unscored 1–3-minute Speaking tasks. For the purposes of this paper, we are only looking at the extended Speaking tasks, as the read aloud task does not require the generation of any original speech, and the unscored tasks are simply intended as speaking samples for test users and do not contribute to the final score. Table 13 presents a comparison of the speaking tasks in the two tests.

Table 13. *Comparison of Speaking tasks on IELTS and DET*

	IELTS	DET
Number of tasks	Three tasks	Two task types; four total tasks
Task description	All three tasks are embedded within an oral interview with an examiner. Task 1: introduction and interview (general questions on familiar topics) Task 2: Long turn (candidate is asked to speak for 1–2 minutes on a topic) Task 3: Discussion (elaborate on issues related to Task 2)	All tasks are computer-based. Picture description (three): Speak for at least 30 seconds about a photo. Question response: Respond orally to a short aural or written prompt. Prompts are graded in difficulty.
Functions elicited	providing personal information, expressing and justifying opinions, explaining, suggesting, speculating, expressing preferences, comparing, summarizing, narrating*	describing, providing personal information, expressing and justifying opinions, explaining, narrating
Timing	11–14 minutes total, including 1 minute of preparation time (Task 2)	20 seconds of preparation time per task; candidates must speak for 30–90 seconds; total speaking time 2–6 minutes

*This list comes from Taylor and Chan's (2015) analysis; other discourse management functions may be possible.

Table 14, from Taylor & Chan (2015), summarizes the levels, outputs, and information sources involved in the cognitive process of speaking. The middle column lists the six stages of cognitive processes in speaking. The conceptualization stage marks the initial generation of an idea the speaker intends to express, which is then processed to form a general framework for the utterance (grammatical encoding) and converted to phonologically realized word strings (morpho-phonological encoding). The speaker then forms a set of neural instructions to the articulators (phonetic encoding) to finally produce the intended utterance (articulation). Finally, at the self-monitoring stage, the speaker evaluates the utterance, which may lead to self-repair. The rightmost column summarizes the outputs of each processing stage, which immediately become the information sources feeding into the next stage of the processing system. Together with the output of a previous processing stage, we see a list of information sources for each stage in the left column of Table 14. A speaker's command of these information sources plays an important role in the quality of their produced utterance.

Table 14. *Cognitive processing model of speaking ability (Taylor & Chan, 2015, adapted from Field, 2011, pp. 74-77)*

Information sources feeding into phases of the processing system	Cognitive processes	Outputs of processing
Speaker's general goals World knowledge Knowledge of listener Knowledge of situation Recall of discourse so far Rhetoric and discourse patterns	**CONCEPTUALISATION**	Pre-verbal message
Recall of ongoing topic Syntax Lexical knowledge Pragmatic knowledge Knowledge of formulaic chunks Combinatorial possibilities (syntactic/ collocational)	**GRAMMATICAL ENCODING:** constructing a syntactic frame forming links to lexical entries	Abstract surface structure
Lexical knowledge Phonological knowledge	**MORPHO-PHONOLOGICAL ENCODING:** conversion to linguistic form	Phonological plan
Syllabary: Knowledge of articulatory settings	**PHONETIC ENCODING:** conversion to instructions to articulators; cues stored in a speech buffer.	Phonetic plan
	ARTICULATION: execution of instructions	Overt speech
Speaker's general goals Target utterance stored in buffer Recall of discourse so far	**SELF-MONITORING**	Self-repair

Field's model is limited in that it only considers the act of producing an utterance, and does not take into account the skills required for spoken interaction, or what is known as interactional competence. Interactional competence is defined by Galaczi and Taylor (2018) as "the ability to co-construct interaction in a purposeful and meaningful way, taking into account sociocultural and pragmatic dimensions of the speech situation and event" (p. 226). While a detailed discussion of interactional competence is beyond the scope of this paper, aspects of interactional competence include topic and turn management, interactive listening, breakdown repair, and non-verbal/visual behaviors, all of which contribute to successful oral communication.

In evaluating the cognitive validity of IELTS and DET Speaking test tasks, we focus on how the cognitive processes employed by the test takers may differ between the two tests. Table 15 summarizes the relevant features for the two tests.

Table 15. *Comparison of cognitive processes in speaking on IELTS and DET*

	IELTS	DET
Cognitive processing: Levels of speaking	**Conceptualization** **Grammatical encoding** **Morpho-phonological encoding** **Phonetic encoding** **Articulation** **Self-monitoring**	(Conceptualization)* **Grammatical encoding** **Morpho-phonological encoding** **Phonetic encoding** **Articulation** (Self-monitoring)
Cognitive processing: Interaction pattern and planning time	Reciprocal (face-to-face) Planning time allowed (1 minute)	Non-reciprocal (computer-delivered) Planning time allowed (20 seconds)

*Parentheses indicate that these processes may be invoked only minimally, given the brevity of the task.
Bold face indicates that these processes are invoked in responding to the task.

While both tests do require spoken production, and thus invoke at some level all of the cognitive processes involved in speaking monologically, it can be argued that the shorter time for preparation and shorter overall speaking time in DET leaves less of an opportunity for either conceptualization or self-monitoring. DET does not require sustained speaking on a topic for more than 30 seconds; while 90 seconds are allowed for the task, the "submit" button is enabled as soon as the 30-second mark is reached, and the DET Guide encourages test takers to "come to a natural conclusion" (p. 28) once the submit button is enabled.

Furthermore, it is clear from Table 15 that the tests are not at all equivalent in terms of their interaction patterns. Crucially, there is no opportunity to demonstrate the ability to interact with another speaker in real time on the test (see also Wagner (2020) for a critique of this aspect of DET).

Table 16 summarizes the relevant contextual factors that affect speaking, i.e., the nature of the topics to be discussed. The IELTS analysis comes from Taylor and Chan (2015); for the DET, the two researchers independently coded the few DET Speaking tasks that are publicly available and then discussed any discrepancies to come to a consensus.

Table 16. *Analysis of context validity variables in speaking on IELTS and DET*

	IELTS	DET
Domain	**Social** Work **Academic**	**Social** Work Academic
Discourse mode	**Descriptive** **Historical/biographical** **Expository** **Argumentative** **Instructive**	**Descriptive** Historical/biographical **Expository** Argumentative Instructive
Content knowledge	General More general than specific **Balanced** More specific than general Specific	General **More general than specific** Balanced More specific than general Specific
Cultural specificity	**Neutral** Mostly neutral Balanced Mostly specific Specific	**Neutral** **Mostly neutral** Balanced Mostly specific Specific
Nature of information	Only concrete **Mostly concrete** **Fairly abstract** Mainly abstract	**Only concrete** **Mostly concrete** Fairly abstract Mainly abstract
Topic familiarity	**Familiar** **Fairly familiar** **Neutral** **Somewhat unfamiliar** **Unfamiliar**	**Familiar** Fairly familiar Neutral Somewhat unfamiliar Unfamiliar
Knowledge of criteria	Band descriptors are made public on website	Grammar, vocabulary, and mechanics emphasized in materials intended for test takers

As can be seen in the table, DET Speaking topics appear to be more familiar, less abstract, and in general less academically oriented than the IELTS Speaking tasks. Particularly at the lower levels, prompts tend to be descriptive in nature (e.g., "describe aloud the image below"; "talk about a hobby or activity that you enjoy"). Combined with the lower cognitive demands of the DET Speaking relative to IELTS, this analysis suggests that IELTS Speaking has greater construct coverage than DET.

Writing

The socio-cognitive framework outlined in Weir (2005) and expanded in Shaw and Weir (2007) describes six main cognitive processes involved in writing, similar to those for speaking outlined earlier. These are the following:

1. Macro-planning: gathering ideas and identifying the task constraints (genre, readership, goals)
2. Organization: ordering ideas, identifying relationships among them, and prioritizing them in terms of their importance to the overall thesis
3. Micro-planning: planning at both the sentence and paragraph level

4. Translation: converting abstract ideas into linguistic form
5. Monitoring: evaluating the text for mechanical accuracy, and at more advanced levels, for adherence to the writer's intention and intended argument structure
6. Revising: making corrections or adjustments to the text as a result of monitoring

The only scored task in DET that elicits any of these cognitive processes are the four extended Writing tasks presented to each candidate (we are ignoring the dictation task for this analysis, as it does not involve the generation of any original content). As a reminder, these tasks include one picture description task and three prompted short responses. Each task has a five-minute limit. In contrast, the IELTS Writing section consists of two longer tasks, for a total of 60 minutes (see comparison in Table 17). Thus, the contextual features of the two tests will determine the degree to which these cognitive processes are evoked; in particular, the shorter tasks in DET presumably offer less scope for macro-planning and revision (see Table 18).

Table 17. *Comparison of Writing tasks*

	IELTS	DET
Number of tasks	Two	Five
Task description	Task 1: Describe or explain information presented in a chart, graph, or table. Task 2: Write an essay in response to a point of view, argument, or problem.	Picture description (3): Write at least one sentence describing a photo. Question response: Write a short response to a question prompt.
Purpose	Task 1: Transfer information from multiple sources to describe, summarise or explain. Task 2: Write a persuasive essay to defend or attack an argument or opinion.	Demonstrate vocabulary and syntactic knowledge. Provide an opinion or personal information.
Timing	60 minutes total Recommended: 20 minutes on Task 1, 40 minutes on Task 2	Five minutes per task, 20 minutes total
Text length of expected response	Task 1: at least 150 words Task 2: at least 250 words	Picture description: at least one sentence Question response: at least 50 words Writing sample: Write for at least three minutes (no minimum word count)
Weighting	Task 2 is weighted twice as much as Task 1	Unclear how writing tasks figure into the final score

Table 18. *Cognitive processing in Writing tasks*

	IELTS	DET
Cognitive processing	**Macroplanning**	Macroplanning
	Organization	Organization
	Microplanning	**Microplanning**
	Translation	**Translation**
	Monitoring	**Monitoring**
	Revising	Revising

The important contextual features include the number of tasks, response format and genre, source texts, domain, topic, purpose, knowledge of criteria, writer-reader relationship, timing, text length, and skills focus. Some of this information is in the description above. A comparison can be found in Table 19.

Table 19. *Analysis of context validity variables in Writing on IELTS and DET*

	IELTS	DET
Domain	**Social** Work **Academic**	**Social** Work **Academic**
Discourse mode	**Descriptive** Historical/biographical **Expository** **Argumentative** Instructive	**Descriptive** Historical/biographical **Expository** **Argumentative** Instructive
Content knowledge	**General** **More general than specific** **Balanced** More specific than general Specific	**General** **More general than specific** **Balanced** More specific than general Specific
Cultural specificity	**Neutral** **Mostly neutral** Balanced Mostly specific Specific	**Neutral** **Mostly neutral** Balanced Mostly specific Specific
Nature of information	**Only concrete** **Mostly concrete** **Fairly abstract** Mainly abstract	**Only concrete** **Mostly concrete** Fairly abstract Mainly abstract
Knowledge of criteria	Band descriptors are made public on website	Grammar, vocabulary, and mechanics emphasized in materials intended for test takers

Taylor and Chan (2015) present data from Banerjee, Franceschina and Smith (2007) that summarize the lexical and syntactic complexity of IELTS Writing responses that are scored at bands 7 and 8 as a way of verifying that the scripts match the band descriptors. For example, lexical variables examined include the percentages of words that fall into the first 1,000 and 2,000 most frequently used words, percentages of words found in the Academic Word List (AWL), type/token ratio, and lexical density. These variables are related to the band descriptor, which states that for band 7, "candidates are expected to use a sufficient range of vocabulary to allow some flexibility and precision and use less common lexical items with some awareness of style and collocation." Similar evidence for syntactic complexity, cohesion, and accuracy is provided.

To our knowledge, Duolingo does not provide examples of scripts from any of their Writing tasks, so it is impossible to provide comparative data. However, given the general nature of the prompts and the brevity of the expected response (at least 50 words, maximum of five minutes), it is unlikely that most responses exceed 150 words[6]. It would therefore be somewhat surprising to find similar levels of lexical and syntactic complexity, as well as a variety of cohesive devices, in DET written responses.

6 Barkaoui (2016) found that L2 English students with high keyboarding skills typed an average of 40 words per minute on a 2-minute typing test, which only includes copying, not composing.

V. Scoring validity

In this section of the report, we discuss the scoring validity of IELTS and DET. We have elected to discuss scoring validity for both tests here, rather than within the discussion of each skill, principally because all items on the DET are scored automatically and there is no easy way to separate out the scoring of distinct skills. As noted earlier, DET provides a total score, along with subscores that combine skills along the axes of oral/written, on the one hand, and reception/production, on the other (see Figure 1 above). In this section, we provide an overview of scoring validity, then discuss the scoring of each test and provide a comparison. Finally, we look at relationships between the two tests.

Scoring validity can be considered a superordinate term for the various aspects of the testing process that can impact the reliability (consistency) of scores (Taylor & Galaczi, 2011, p. 171). Quoting Shaw and Weir (2007), Taylor and Galaczi further state that scoring validity "accounts for the extent to which test scores are based on appropriate criteria, exhibit consensual agreement in marking, are as free as possible from measurement error, stable over time, consistent in terms of content sampling and engender confidence as reliable decision-making indicators" (2007:143).

For speaking and writing, which are typically evaluated by human raters using a rating scale, relevant aspects of scoring validity include the following (Taylor & Galaczi, 2011):

- Criteria/rating scale
- Rating process
- Rating conditions
- Rater characteristics
- Rater training
- Post-exam adjustments
- Grading and awarding.

While all these elements are important, perhaps the most critical factor in scoring productive items is the degree to which independent raters agree with each other. A variety of inter-rater reliability statistics can be reported, including a simple correlation coefficient, a Kappa coefficient, or the percentage of cases in which raters agree on the exact score (e.g., both raters give a 7 out of 9) or an adjacent score (e.g., one rater scores 6 and another scores 7, with the reported score being the average of the two scores). Other statistical methods for investigating inter-rater reliability include Generalizability Theory (e.g., Brennan, 1992; Huang, 2012) and Many-Facet Rasch measurement (e.g., McNamara, Fan, Knoch & Rossner, 2019).

For listening and reading, which are typically assessed using item types that can be scored correct/incorrect, important considerations include the following (Khalifa & Weir, 2009):

- Item difficulty
- Item discrimination
- Internal consistency
- Error of measurement
- Marker reliability
- Grading and awarding.

Overall test reliability is an essential component of a test. For objectively scored items, an internal consistency coefficient is typically reported, indicating the degree to which test items are functioning in a similar fashion (Popham, 1990, p. 55). Conceptually similar, but

more difficult to obtain in real-life situations, are test-retest reliability, and alternate forms reliability. Test-retest reliability refers to the situation where candidates take the same test at two different times, while alternate forms reliability is calculated when different forms of the test are administered to the same population. In all cases, what is reported is the equivalent of a correlation coefficient, with values closer to 1 indicating higher reliability.

In Tables 20 through 22, we present a comparison of scoring across the four skills. For IELTS, scoring is done separately by section. For DET, Writing and Speaking scores are based on extended Writing and extended Speaking, respectively. For reading and listening, we present a single table since the considerations are similar for both tests, and DET does not report by section. Note that the reliability statistics (test-retest and internal consistency) are reported for production in Tables 20 and 21 and for the other subscores in Table 22.

Table 20. *Writing scoring validity*

	IELTS	DET
Raters	Trained examiners	Automated scoring
Scoring approach	Analytical; four separate scores are generated for each task.	Based on machine learning algorithm; single score generated for all writing items together
Number of raters	Single rater; double rated under some circumstances	Single computer-generated score
Setting of scoring	Tests are scored at test centers worldwide and monitored centrally	No information; presumably scored by computers housed on Duolingo's campus
Scoring criteria	Task achievement/response Coherence and cohesion Lexical resource Grammatical range and accuracy	Grammatical accuracy Grammatical complexity Lexical sophistication Lexical diversity Task relevance Length
Score reporting	Skill scores are reported as whole or half bands from 0–9	Scores are not reported separately but combined with other scores to calculate overall scores along with relevant subscores (literacy, production)
Reliability	Generalisability coefficients based on examiner certification data: .81–.89	Machine-human agreement:* † Human:Human κ =.68 Human:Machine κ = .82 Human:Machine κxv = .73 Internal consistency: Production: .75† Test-retest reliability: Production: .88 SEM: Production: 7.74 ~~10.85~~

*Kappa (κ) is a measure of the probability of agreement of scores with chance agreement factored out. κxv represents the agreement when 10-fold cross validation is used; that is, ten different combinations of training and testing responses.

† Not reported in latest DET Manual. Statistics from LaFlair & Settles (2020).

Source: IELTS performance statistics can be found at https://www.ielts.org/for-researchers/test-statistics; Unless otherwise specified, DET statistics are from DET Manual.

‡Updated in 2022 manual: Test-retest reliability – Production: .88
§Updated in 2022 manual: SEM – Production: 7.74

Source: IELTS performance statistics can be found at https://www.ielts.org/for-researchers/test-statistics; DET statistics are from the DET Manual.

Table 21. *Speaking scoring validity*

	IELTS	DET
Raters	Trained examiners	Automated scoring
Scoring approach	Analytical; four separate scores are generated for each task	Based on machine learning algorithm; single score generated for all speaking items together
Number of raters	Single rater; double-rated under some circumstances	Single computer-generated score
Setting of scoring	Tests are scored at test centers worldwide and monitored centrally	No information; presumably scored by computers housed on Duolingo's campus
Scoring criteria	Fluency and coherence Lexical resource Grammatical range and accuracy Pronunciation	Grammatical accuracy Grammatical complexity Lexical sophistication Lexical diversity Task relevance Length Fluency & acoustic features
Score reporting	Skill scores are reported as whole or half bands from 0–9	Scores are not reported separately but combined with other scores to calculate overall scores along with relevant subscores (conversation, production)
Reliability	Generalizability coefficients based on examiner certification data: .83–.86	Machine-human agreement:* † Human:Human κ = .77 Human:Machine κ = .79 Human:Machine κxv = .77 Internal consistency: † Production: .75 Test-retest reliability Production: .81 SEM: Production: 7.74

*Kappa (κ) is a measure of the probability of agreement of scores with chance agreement factored out. κxv represents the agreement when 10-fold cross-validation is used; that is, ten different combinations of training and testing responses.

† Not reported in latest DET Manual. Statistics from LaFlair & Settles (2020).

Source: IELTS performance statistics can be found at https://www.ielts.org/for-researchers/test-statistics; DET statistics are from the DET Manual.

The tables above must be interpreted in light of the differences in task. As a reminder, IELTS Writing and Speaking scores are based on much longer stretches of discourse produced by test-takers, while DET tasks are much shorter and more constrained. IELTS scores, being produced by single raters, may come in for some criticism in terms of reliability, but the reported generalizability statistics for IELTS are somewhat higher than the various reliability measures reported by Duolingo, though the statistics are not directly comparable. The differences between aspects of texts that are salient to human raters and those that can be measured automatically have been pointed out by numerous scholars (see Deane, 2013 for a summary); at best, as even Duolingo admits, those features of a text that can be measured can only serve as a proxy for factors that are important to human raters. As Deane (2013, p. 18) states, "if the focus of the assessment is to quality of argumentation, sensitivity to audience, and other such elements to differentiate among students who have already achieved fundamental control of text production processes, the case for automated essay scoring is relatively weak.'

As for listening and reading, the reliability indices in Table 22 suggest that both tests are sufficiently reliable in terms of internal consistency. IELTS has no data for test-retest reliability, so it is not possible to make direct comparisons of the tests in this area.

Table 22. *Listening and reading scoring validity**

	IELTS	DET
Scoring approach	Scanned answer sheets for dichotomous items; trained raters using a mark scheme for section 2	Automated scoring
Weighting	All items equally weighted	Weighted averages are calculated for each CAT item type and are used to create a total score and subscores. Manual does not clearly say how speaking and writing tasks are factored into scores.
Reliability	Listening: Average Alpha across 16 versions (2020 data): .92 Reading: Average Alpha across 16 versions (2020 data): .90	Test-retest Literacy: .80 Conversation: .78 Comprehension: .76 Total: .82 Internal consistency: Literacy: .88 Conversation: .93 Comprehension: .95 Total: .95
Standard Error of Measurement	Listening: .37 (in terms of score bands) Reading: .40	Literacy: 6.48 Conversation: 5.67 Comprehension: 4.12 Total: 3.92

*For DET, calculation of total scores and subscores also incorporates extended Speaking and Writing tasks, so these cannot be completely separated out.

Source: IELTS performance statistics can be found at https://www.ielts.org/for-researchers/test-statistics; DET statistics are from the DET Manual.

VI. Criterion-related validity

Criterion-related validity has to do with the relationship between one test and another of the same ability, and with the ability of a test to predict future performance. In this section of the report, we discuss how the two tests relate to each other and to the CEFR.

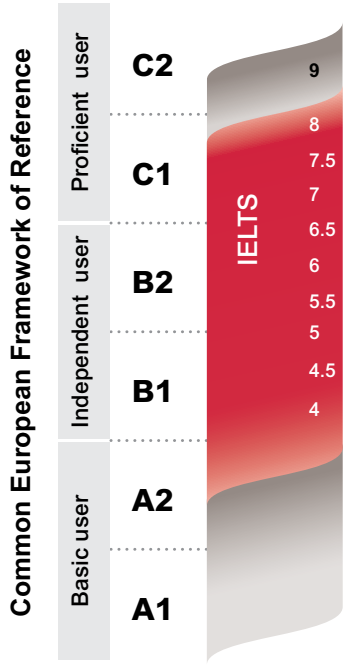
Duolingo provided a concordance table between IELTS and DET, partially replicated below in Table 23, based on the performance of 991 test takers who took both tests. Each IELTS band is associated with two score points (10 total points) on DET. The correlation between the two tests, based on 991 test takers, is .78, suggesting that scores on the two tests have a moderate to strong relationship. Correlations between scores on the Writing and Speaking sections of the two tests are lower, at .42 and .54, respectively, showing a weaker relationship. IELTS has not produced similar research, nor has any independent researcher conducted a study comparing the two tests, to the best of our knowledge.

In its concordance table, Duolingo also includes descriptors from the CEFR, though the DET Manual does not give any indication that the recommended procedures for aligning the test to the CEFR were followed by Duolingo (see Figueras, North, Takala, Verhelst, & Van Avermaet, 2005). IELTS, in contrast, has conducted numerous studies exploring the relationship of band scores to the CEFR. Figure 8 (https://www.ielts.org/-/media/pdfs/comparing-ielts-and-cefr.ashx?la=en) shows the comparison of IELTS scores and the CEFR levels.

Table 23. *Concordance table between IELTS and DET (Source: Duolingo)*

IELTS	9	8.5	8	7.5	7	6.5	6	5.5
DET	155–160	145–150	135–140	125–130	115–120	105–110	95–100	85–90
Description (CEFR)	Advanced (120–160) • Can understand a variety of demanding written and spoken language including some specialized language use situations. • Can grasp implicit, figurative, pragmatic, and idiomatic language. • Can use language flexibly and effectively for most social, academic, and professional purposes.				Upper intermediate (90–120) • Can fulfill most communication goals, even on unfamiliar topics. • Can understand the main ideas of both concrete and abstract writing. • Can interact with proficient users fairly easily.			

Figure 8. *Alignment of IELTS scores with the CEFR scale*



While IELTS uses CEFR terminology, distinguishing among basic, independent, and proficient users, Duolingo uses terms that may be more familiar to Americans (advanced; upper intermediate). Additionally, IELTS references the CEFR Can Do statements directly in their literature, but it is not immediately clear what process was used to modify the CEFR statements for the DET or to map them on to scores. Interestingly, while IELTS research suggests that band 7 represents Level C1, the Duolingo table implies that this level is still considered "upper intermediate". For these reasons, the CEFR levels provided by Duolingo are to be interpreted with caution.

Figure 16 of the DET Technical Manual (p29) shows a scatterplot comparing scores on IELTS and DET. The orange line represents the regression line, which can be interpreted as the predicted score on one test given the score on the other. Points on the graph to the left of the line represent cases in which test takers received higher scores than

predicted on IELTS than DET, and points to the right represent test takers scoring higher than predicted on DET than IELTS. A close inspection of the scatterplot reveals that there are more test takers scoring higher on DET than IELTS between IELTS bands 4 and 5.5, while more test-takers score higher on IELTS than DET from bands 7 and higher. The area between 5.5 and 7, which is typically the range of scores where high-stakes decisions are made, shows the widest variability between scores on the two tests, suggesting that the relationship between the two tests may not be as straightforward as Duolingo implies.

VII. Consequential validity

A thorough investigation of test consequences is beyond the scope of this paper. However, one way to gain insight into how tests affect teaching and learning is to examine what test-takers themselves say about the tests. Since China is a major market for both tests, we collected information about test taker perceptions of both tests from 10 Chinese online discussion platforms (14 posts in total) between 2020 and 2021.[7] These sites were chosen because they are popular among test takers of both tests to communicate their test preparation strategies as well as their unfiltered opinions about the tests. All 14 posts chosen for this project compared DET with IELTS. While this is a small sample, it does provide some insights into the potential washback of the tests in terms of how test-takers prepare for the test.

Perhaps unsurprisingly, test-takers do not seem to discuss the validity of the tests. Most online discussions of the comparison between IELTS and DET focus on three main areas, as shown in Table 24.

Table 24. *Online posts comparing IELTS with DET*

Discussion topic	Mentioned in posts
Test difficulty	N = 11 (1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 14)
Test accessibility	N = 10 (1, 2, 3, 4, 5, 7, 8, 9, 11, 12)
Test preparation	N = 7 (1, 2, 8, 9, 12, 13, 14)

Test difficulty

The most frequently discussed comparative aspect of the two tests was the relative difficulty, as reflected in the scores. Most posters who took both tests within a short amount of time reported that the score they received on DET was higher than their IELTS score. For example, one poster stated that she scored a 7 on IELTS and 125 on DET (post 1). Another post indicated that a test taker who "received a score of 5 on IELTS was able to score 120 on DET" (post 10, translated by author 2). Note that, according to the concordance reported above (Table 23), these DET scores should correspond to a higher IELTS score.

The impression that DET is easier than IELTS appears to be shared by at least some institutions. Two posters reported that the schools they applied to advised them to take the DET after receiving a score on IELTS, implying that they would more easily reach the required minimum score on DET. Speaking is frequently perceived as a skill that is easier

7 These posts were collected before the addition of Interactive Reading to the DET.

on DET than IELTS. One poster wrote: “the DET provides just as much instruction for the Speaking tasks as IELTS and TOEFL, but requires much less input from the test takers” (post 3, translated by author 2).

Test accessibility

Not surprisingly, since many posters took or were considering the DET as an alternative to the conventionally in-person IELTS test, many posts focus on the comparison of test accessibility. These discussions include comparisons of cost, test length, ease of understanding of the online interface, and the reliability of the technology.

A majority of posters who commented on this aspect of the test favor the DET because of its much lower registration fee, shorter test length, and shorter wait time for the score reports. In terms of cost, many posters reported taking DET multiple times due to its lower registration fee and the convenience of taking it from home. In addition, all posts mention the fact that the online format of the test is much less time and labor consuming than the in-person test. On the other hand, several posts mentioned the downside of the DET, specifically on the reliability of the technology. For example, post 2 mentioned that some test takers failed the test because they were wearing jewelry that was identified as a cheating device. A few posts (e.g., post 2 and 5) also mentioned the difficulty in staying still throughout the testing process, since movement can be identified as cheating.

Test preparation

Posters who commented on this aspect of the test indicate that preparation for IELTS involves rigorous drilling of the tasks and analysis of question patterns, but this strategy is not applicable to preparing for the DET. Some posters felt that DET is not as coachable as IELTS and thus required more knowledge of English words and structures. In one post (post 10), a test taker elaborated on their impression of the two tests:

“In my opinion, IELTS and TOEFL are more rigid and focus more on academic content. But there are more strategies we could use for IELTS and TOEFL. For example, if you do not understand the word in question, you could find the answer by searching for other key words or browsing the context. However, DET is more realistic and flexible. Its question types are more varied, and it requires test takers to react fast. We don’t have test-taking strategies that help us to answer those questions. If we know the language, we do well. If we don’t, we don’t.”

It seems that the variety of task types and the lack of available test-taking strategies for DET lead some test takers to believe that DET is more relevant to testing language ability, as opposed to test-taking strategies. For example, post 12 stated that “comparing to IELTS, DET has more task types that are not familiar to test takers. It is more difficult to figure out the question patterns, and it requires more solid foundation for language use rather than ‘techniques’ for test taking.” Another post (post 2) echoes this point by emphasizing the importance of vocabulary, listening comprehension, and pronunciation to achieving high scores.

Indeed, in terms of test-taking strategies, since IELTS has been around for decades, test takers have ready access to a plethora of resources on practice tests and test-taking strategies. On the contrary, DET is relatively new and has just started to be used for admission purposes, so not many test takers are familiar with its test format. The only task for which posters had suggestions for preparation was the C-test. One poster (post 5) advises the following strategy:

1. read the first and last sentence first
2. fill in the blanks in turn
3. pay attention to the structure of sentences and clauses
4. identify the part of speech of the target words
5. use semantic category to narrow down possible word options

There were no suggestions for preparing for other task types, such as the aural/visual yes-no questions, the dictation task, the elicited imitation task, or the extended Speaking and Writing tasks. However, given the high-stakes nature of admissions testing, it does not seem far-fetched to predict that test preparation schools may soon provide "rigorous drilling" of words vs. non-words, single-sentence dictation, and 30-second oral picture descriptions, to the detriment of practicing essential academic skills such as guessing words in context from a reading passage, listening for key words in a lecture, or writing a well-developed essay, which are strategies often mentioned by test takers preparing for IELTS.

VIII. Summary and conclusion

In this report, we have provided an in-depth comparison of IELTS and DET in terms of the factors that are important for test users to consider when deciding whether a test is appropriate for a given purpose. Our analysis demonstrates that, compared to IELTS, DET test tasks under-represent the construct of academic language proficiency as it is commonly understood, i.e., the ability to speak, listen, read, and write in academic contexts. Most of the DET test tasks are heavily weighted towards vocabulary knowledge and syntactic parsing rather than comprehension or production of extended discourse, though the recent addition of Interactive Reading addresses this lack somewhat. Scores on the two tests are correlated, which might suggest that DET is a reasonable substitute for IELTS, given its accessibility and low cost. Of course, knowledge of lexis and grammar are essential enabling skills for higher-order cognitive skills, and a test that focuses on these lower-level skills can be useful for making broad distinctions between low, intermediate, and high proficiency learners. However, potential test users should be aware of the limitations of DET in terms of predicting academic success.

It may be useful to recall that, some 20 years ago, another well-known large-scale English proficiency test, the TOEFL, underwent a complete overhaul to focus less on the enabling skills of grammar and vocabulary and to emphasize longer, more authentic academic Speaking and Writing tasks. This revision was undertaken in part because ESL and EFL teachers felt that the discrete test tasks, while highly reliable, were not relevant to the language needs of their students, and in part because test users found that students with high scores "arrive[d] on campus with insufficient writing and oral communication skills to participate fully in academic programs" (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000). We note that DET has already begun to modify its content with the addition of the Interactive Reading section and a second scored writing task, perhaps in response to similar pressures. It remains to be seen whether a test that relies primarily on the efficiencies of machine learning and natural language processing at the expense of cognitive and context validity can escape the same fate.

References

Banerjee, J., Franceschina, F., & Smith, A. M. (2007). *Documenting features of written language production typical at different IELTS band score levels*. International English Language Testing System (IELTS) Research Reports Volume 7. Retrieved from: www.ielts.org/-/media/research-reports/ielts_rr_volume07_report5.ashx

Barkaoui, K. (2016). What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal*, *100*(1), 320–340.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34.

Cardwell, R., LaFlair, G. T., & Settles, B. (2022). *Duolingo English test: Technical Manual*. Retrieved from: duolingo-papers.s3.amazonaws.com/other/det-technical-manual-current.pdf

Chalhoub-Deville, M., & O'Sullivan, B. (2021). *Validity: Theoretical Development and Integrated Arguments.* Bristol, CT: Equinox Publishing.

Coxhead, A. (2000). *Academic Word List*. Retrieved from: https://www.wgtn.ac.nz/lals/resources/academicwordlist

Cutler, A., & Clifton, C. (1999). Comprehending spoken language: A blueprint of the listener. In C. M. Brown and P. Hagoort (eds.), *The Neurocognition of Language* (pp. 123–166). Oxford: Oxford University Press.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24.

Duolingo, Inc. (2021). *The Duolingo English Test Official Guide*. Retrieved from: englishtest.duolingo.com/guide

Field, J. (2009). *Listening in the Language Classroom*. Cambridge: Cambridge University Press.

Field, J. (2011). Cognitive validity. In L. Taylor (ed.), *Examining Speaking: Research and Practice in Assessing Second Language Speaking* (pp. 64–111). Cambridge: UCLES/Cambridge University Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (eds.), *Examining Listening: Research and Practice in Assessing Second Language Listening* (pp. 77–151). Studies in Language Testing volume 35. Cambridge: UCLES/Cambridge University Press.

Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing, 22*(3), 261–279.

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219–236.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, *17*(3), 123–139.

Hunt, K. W. (1964). *Differences in grammatical structures written at three grade levels, the structures to be analyzed by transformational methods.* Report No. CRP-1998. Tallahassee: ERIC Document Reproduction Service No. ED 003 322, Florida State University.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 Framework*. Princeton, NJ: Educational Testing Service.

Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review*, *64*(1), 215–238.

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading.* Studies in Language Testing volume 29. Cambridge: UCLES/Cambridge University Press.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*(1), 47–84.

LaFlair, G. T. (2020). *Duolingo English Test: Subscores*. Duolingo Research Report DRR-20-03. Duolingo, Inc. Retrieved from https://duolingo-papers.s3.amazonaws.com/reports/subscore-whitepaper.pdf

LaFlair, G. T., & Settles, B. (2020). *Duolingo English test: Technical Manual*. Retrieved from: duolingo-papers.s3.amazonaws.com/other/det-technical-manual-current.pdf

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

McKay, T. (2019). *More on the Validity and Reliability of C-Test Scores: A Meta-Analysis of C-Test Studies*. PhD dissertation, Georgetown University.

McNamara, T., Knoch, U., Fan, J., & Rossner, R. (2019). *Fairness, Justice & Language Assessment*. Oxford University Press.

Pearson, W. S. (2019). Critical perspectives on the IELTS test. *ELT Journal, 73*(2), 197–206.

Pilcher, N., & Richards, K. (2017). Challenging the power invested in the International English Language Testing System (IELTS): Why determining 'English' preparedness needs to be undertaken within the subject context. *Power and Education, 9*(1), 3–17. https://doi.org/10.1177/1757743817691995

Popham, W. J. (1990). *Modern Education Measurement: A Practitioner's Perspective.* Boston: Allyn and Bacon.

Raatz, U. and Klein-Braley, C. 1981: The C-Test - a modification of the cloze procedure. In Culhane, T., Klein-Braley, C. and Stevenson, D.K., editors, Practice and problems in language testing, University of Essex Department of Language and Linguistics Occasional Papers No. 26, Colchester: University of Essex.

Settles, B., T LaFlair, G., & Hagiwara, M. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247–263.

Shaw, S. D., & Weir, C. J. (2007). *Examining Writing: Research and Practice in Assessing Second Language Writing.* Studies in Language Testing volume 26. Cambridge: UCLES/Cambridge University Press.

Taylor, L., & Chan, S. (2015). *IELTS Equivalence Research Project (GMC 133)*. Retrieved from: www.gmc-uk.org/-/media/documents/GMC_Final_Report___Main_report__extended____Final___13May2015.pdf_63506590.pdf

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (ed.) *Examining Speaking: Research and Practice in Assessing Second Language Speaking* (pp. 171–233). Studies in Language Testing volume 30. Cambridge: UCLES/Cambridge University Press.

Wagner, E. (2020). Duolingo English Test, Revised Version July 2019. *Language Assessment Quarterly, 17*(3), 300–315.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.

Appendix. Forum posts comparing IELTS and DET

Post ID	Site	Link to original post
#1	Baidu Posts	https://tieba.baidu.com/p/7529724950?pid=141198522990&cid=0#141198522990
#2	Bilibili	https://www.bilibili.com/read/cv6966089/
#3	BaiduZhidao	https://zhidao.baidu.com/question/1804527966788620227
#4	ChaseDream Forum	https://forum.chasedream.com/forum.php?mod=viewthread&tid
#5	Fox IELTS	http://www.foxiielts.com/special/news?id=ab79cbdf142242cdbcb6b530d28a8b6c
#6	51 Offer	https://www.51offer.com/article/detail_98007.html
#7	Sohu Forum	https://www.sohu.com/a/218900942_100002843
#8		https://www.sohu.com/a/391935073_99918349
#9	5HLX	http://www.5hlx.com/liuxuezixun/3995.html
#10	XinHangdao	https://www.xhd.cn/ielts/zonghe/156983.html
#11	Zhihu	https://zhuanlan.zhihu.com/p/111931531
#12		https://zhuanlan.zhihu.com/p/142319865?ivk_sa=1024320u
#13		https://zhuanlan.zhihu.com/p/147071142
#14		https://zhuanlan.zhihu.com/p/373475309

Accessibility statement.

IELTS is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, please submit our contact form at ielts.org/enquiry and we will respond within 15 working days.