Language Learning & Technology
ISSN 1094-3501
June 2021, Volume 25, Issue 2
pp. 1–25

EMERGING TECHNOLOGIES

Multilevel language tests: Walking into the land of the unexplored

Jesús García Laborda, *Universidad de Alcalá*
Miguel Fernández Álvarez, *Universidad Politécnica de Madrid*

Abstract

This paper compares and analyzes a selection of popular multilevel tests used for quick accreditation of English as a foreign language worldwide. The paper begins by stating the current need of accreditation of English language competence for both academic and professional matters. It then looks at their defining features and differences. After, the different pros and cons are analyzed looking especially at the need to diversify item types since the authors consider that even the most novel tests have a traditional construct that dates back many years. It also proposes new types of items. The paper concludes that a revision of the concept of language construct is necessary considering the specific uses of the language in the 21st century.

Keywords: *Multilevel Tests, Language Tests, Test Construct, Test Differences*

Language(s) Learned in This Study: *English*

APA Citation: García Laborda, J., & Fernández Álvarez, M. (2021). Multilevel Language tests: Walking into the land of the unexplored. *Language Learning & Technology*, *25*(2), 1–25. http://hdl.handle.net/10125/73428

Introduction

Reaching a certain level of language proficiency and certifying competency in a second language has become essential in today's fast-paced global society, especially in the case of English. Whether for educational (such as access to universities or graduation requirements), labor, immigration, or one amongst many other reasons, at some point in our lives we are faced with the need of demonstrating high proficiency in English. There is an array of options in the language testing field that users can choose from, and international language testing companies are constantly advancing and adapting their tests to meet the needs of a rapidly changing world. They have been looking at medium and high stakes multilevel assessments that can serve to provide competence information within just a few hours or days for a variety of purposes as well as reducing the delivery costs, improving the security both in the design and delivery costs, and so on. While multilevel assessments have existed for many years (Roever, 2001), the advances in technology in language testing have facilitated their development and popularization and thus developed significantly the way students are assessed today.

The growth of Computer-Assisted Language Testing (CALT) and Web-Based Language Testing (WBLT) as fields a few decades ago revolutionized the development of language tests, making the paper and pencil test in some cases an alternative among several options available (García Laborda, 2007; Fernández Álvarez, 2016). The role of technology has become an essential component of testing practices, and computers have been increasingly used in medium-stakes (such as the graduation requirements or access to some courses) or high-stakes tests (such as a prerequisite to study in reputed universities abroad) (Shin, 2012; Long et al., 2018). That has been the case recently with the new situation created by the COVID-19 world pandemic, when users were in the need of certifying their language proficiency level for access or graduation purposes, but the options were limited. Most users looked for tests that could be done at home, but concerns about their validity, delivery, security, proctoring, fit of students' equipment or whether test

results would be accepted (consequential validity) by institutions were raised when determining what test to take.

Multilevel high-stakes language computer-based tests such as the International English Language Testing System (IELTS) or the Test of English as a Foreign Language Internet-based Test (TOEFL® iBT) were two of the main alternatives, as they have received significant interest from the research community and their use and validity are in constant revision. However, there are also a number of other tests that have not reached that level and are considered 'medium language tests.' A common characteristic of those tests is the limited amount of research that most of them have received, which is often based on the experience acquired through their use rather than from internal validation studies. With a mid or low stakes impact, they tend to resemble and be based on the research, developments and item bank of larger certification tests such as IELTS (for APTIS) or the Cambridge Suite (for Linguaskill). In general, they are usually adaptive and shorter in time and length, since often the number of items per skill is smaller than in certification tests. They are also delivered flexibly in more locations (including schools, universities, local academies, and so on) and, most importantly, they are always computer-based. The easiness of use and their relative immediacy of delivery make them very appealing both for the students and the institutions that require them. This paper intends to look at some of the most common language proficiency tests, indicating their features and application.

Background

Computer-Assisted Language Testing (CALT) and Web-Based Language Testing (WBLT) use Internet for development and delivery. Web based has become generalized in the last twenty years, but there has been a special need for online testing worldwide during the 2020 crisis due to the COVID-19 pandemic. The situation created by educational institutions' and testing centers' lockdowns have had a strong impact on teaching and testing, with many test takers facing the need to find either accreditation or certification tests that they can take from home. As a consequence, some testing companies have revised or even developed versions of their own tests but with remote proctoring. That has been the case of the TOEFL® iBT Special Home Edition, the TOEIC® Special Home Edition or Linguaskill. Overall, the field of language testing has also seen the need for test reviews like the present study or the one conducted by Isbell and Kremmel (2020), where they present different options for at-home language proficiency tests.

Like other online tests, multilevel tests can be delivered individually both in academic settings but lately, even more importantly, some can be delivered assisted by distance proctoring at the test candidate's place. This has made them ideal to be taken at home and thus fulfilling the requirements of many institutions everywhere. In this specific context, testing organizations can easily modify and adapt to specific conditions and different types of language such as ESP, LSP, Young Learners, etc.

WBLT uses the Internet as a platform for test development and delivery; test input and questions are written in the HTML located on a server and test takers respond to the test items using web browsers such as Internet Explorer, Firefox, Chrome, or Safari (Shin, 2012). Recently, WBLT has been embraced more by language researchers and teachers as a teaching and testing tool because it has the potential to greatly enhance logistical efficiency and flexibility (Ockey, 2009). Test developers can easily upload and update test contents, and test takers can take the test at the place and time of their convenience. Test takers' responses on the test are scored immediately, and scores are reported to all stakeholders more quickly. Various item and test score statistics are available on demand, providing useful information for test developers and users to interpret test scores and revise the test when necessary.

Additionally, WBLT has been known to lead to improved test measurement qualities including reliability and validity (Chapelle & Douglas, 2006). A large number of test takers' responses on true/false and multiple-choice formats can be instantly scored without any errors. Even productive responses are scored

consistently once reliable scoring algorithms are developed and applied to test takers' responses (Bernstein et al., 2010; Carr & Xi, 2010). Further, inter- and intra-rater reliability in assessing test takers' written or spoken responses are not a concern in WBLT using an automated scoring system (Williamson et al., 2004).

Authenticity can also be enhanced because various test formats are possible, including interactive and dynamic features of test input and question types (Chapelle & Douglas, 2006; Huff & Sireci, 2001). For example, computer technology makes it possible to include visual input for online listening tests, more closely reflecting language use in real-world tasks (Ockey, 2007; Wagner, 2010). Thus, WBLT is becoming more widely used in many high-stakes standardized language proficiency exams, such as the TOEFL® iBT and the Pearson Test of English Academic (PTE Academic), as well as in placement and screening tests used for medium-stakes decisions made in FL programs (Bardovi-Harlig & Shin, 2014; Elder & Randow, 2008).

However, there are several aspects that need much more attention and research, such as proctoring, security, identity and authentication (Fernández Álvarez, 2016). Technology has advanced fast, and many online tests nowadays use either live remote proctoring through the use of a video camera or systems that record the testing session and recognize the test takers' actions conducted by AI algorithms that analyze "characteristics of the test performance in order to identify potential indicators of rule breaking and malicious behavior" (Duolingo, Inc., 2020). Furthermore, more than technological concerns, there are still ethical and legal questions (Lowman, 2017) that need to be addressed.

High Stakes Certification Tests

As mentioned earlier, there are two main multilevel tests that are probably the most commonly used worldwide, the IELTS (https://www.ielts.org/) and the TOEFL® iBT (https://www.ets.org/toefl). Since their use and validation, as well as their research, are sound, only a brief overview of each test will be provided in this section.

International English Language Testing System (IELTS)

The IELTS exam is a test that is administered by the University of Cambridge, the British Council and the Australian IDP, and is the most demanded test in the world for both work and educational reasons. It is accepted in more than 9,000 organizations in 140 countries around the world, including companies and academic and government institutions. It is the only English test that is accepted in all countries that require a language test for immigrants.

The exam is scored on a scale of 1 to 9. Test takers receive a score for each part as well as an average score. To access a postgraduate or Master's degree at a university in Australia, England or New Zealand, for instance, a minimum score of 6.5 is required, with neither part receiving a score below 6.0. With a price of €210, the IELTS exam has four parts (Listening, Reading, Writing and Speaking), and the total duration of the exam is 2 hours and 45 minutes. Provisional results are provided online 13 days after completing the exam. English certification through IELTS is valid for two years.

TOEFL® iBT

The TOEFL® iBT, developed by Educational Testing Service (ETS), is a test that assesses academic English, which makes it one of the most popular exams used for university entry. While the exam has four different sections (Reading, Listening, Speaking and Writing), tasks are integrative and combine the four skills. It is used worldwide and accepted by more than 11,000 universities in over 150 countries.
The exam has a duration of 3 hours, and fees vary by testing location, with an average price of €250 or $225 USD. Candidates receive a score from 0-30 in each section and an average score from 0-120. Reading and Listening have four proficiency levels (advanced, high-intermediate, low-intermediate, and below low-intermediate). Speaking and Writing include five proficiency levels (advanced, high-intermediate, low-

intermediate, basic, and below basic). Minimum TOEFL® requirements for universities vary. Some of the most prestigious institutions require high 100+ scores, while others accept scores in the 80s or 90s. Scores are valid for two years from the test date.

The latest version of the test is the TOEFL® iBT Special Home Edition, which allows candidates to do the exam from any computer without the need to go to a testing center. With content and format similar to other versions of the exam, the TOEFL® iBT Special Home Edition is proctored online through a system called ProctorU®.

Mid and Low Impact Accreditation Tests

In this section, a description of the tests chosen for this paper (including some general information, test structure, price and, when possible, some research about the tests) is provided. The information presented for each test is up to date as of August 2020, as prices and test structure are constantly under revision and they may change.

One of the criteria for the selection of the tests is their delivery method, as all of them are either computer-based or computer-adaptive tests. Also, as mentioned earlier, these tests have not received as significant international interest as the IELTS or TOEFL® iBT. However, their use is quite popular especially in some specific countries.

The tests are presented in alphabetical order, as follows: Aptis, Duolingo English Test (DET), LanguageCert Test International ESOL, Linguaskill, Pearson Test of English (PTE) Academic, Oxford Test of English, TOEIC® Exams, Trinity's Integrated Skills in English (ISE) exams. At the end of the section there are four tables with summaries about the main aspects previously described for each test. Table 1 focuses on APTIS and Duolingo. Table 2 summarizes the main characteristics of LanguageCert and Linguaskill. In Table 3, we can find details about Pearson Test of English Academic and the Oxford Test of English. Finally, Table 4 provides a summary of the TOEIC® Exams and Trinity's Integrated Skills in English (ISE) exams.

Appendix A also includes tables with a description of the structure of each test, indicating the number and types of tasks for each component. Descriptions vary depending on the information provided in the test specifications, which in some cases are more detailed than others.

APTIS

It is developed by the British Council and is recognized by many organizations and universities. Its drawback is that it is only valid in Spain, and like the Cambridge exams, it does not expire. The APTIS is the "great unknown" amongst current English certifications, not for what it can contribute and its recognition, but as one of the latest additions to the list of exams for accreditation of English language proficiency.

Test takers receive numerical scores from 0 to 50 both for the grammar and vocabulary sections and for each of the language skills (to which the corresponding level in the CEFR – Common European Framework of Reference for Languages is added) and an overall grade in accordance with the Common European Framework. Results are provided within 48 hours and everything is done by computer in a testing center. In the APTIS exam, test takers can choose which parts to certify. Assuming they want everything certified, its price is €64. With a length of approximately 3 hours (depending on the version), the APTIS exam consists of the following parts: Grammar and Vocabulary, Reading, Listening, Speaking and Writing. Candidates can choose between three exams: General (A1-C), Advanced (level B1-C2) and for Teens (level A1-C).

Much of what has been researched about the APTIS test has been funded by the British Council through the Assessment Research Grant program, as evidenced in the number of projects listed on their website. One of their goals is the validation of APTIS and other British Council assessment projects, which studies like Tavakoli et al. (2017) focus on. There are studies that present some concerns about the exam. Knoch et al. (2016), for their part, are worried about how raters are trained, and they make recommendations for a more controlled rating process in the Speaking component. They are particularly concerned with the "level of online support provided" (p. 105). Also, in their study about the effect of response order on candidate viewing behavior and item difficulty, Holzknecht et al. (2020) claim that "the results suggest that the spatial location of the key in MC listening tests affects the amount of processing it receives and the item's difficulty" (p. 2). Impact studies about the effects of anxiety on test takers while doing the APTIS test also conclude that more concrete questions are needed in order to minimize this effect (Valencia Robles, 2017).

Duolingo English Test (DET)

Duolingo English Test (DET), developed by Duolingo Inc., is a computer-adaptive English proficiency test that can be taken anywhere without the need to do it from a testing center. One of the main characteristics of this test is that there is no need to schedule or make any appointments, which makes it very convenient for test takers. Results, which range from 10-160 points, are usually provided within two days, and they are accepted in over 2000 institutions worldwide, especially for higher education admission in English speaking countries. The registration fee is $49 USD.

As opposed to other tests, the Duolingo English Test is not organized by the four skills of listening, reading, speaking and writing. Instead, tasks are integrative, and they focus on Literacy (reading & writing), Comprehension (reading & listening), Conversation (listening & speaking) and Production (writing & speaking). The first part of the test is adaptive and includes the following item types: c-test, audio yes/no vocabulary, visual yes/no vocabulary, dictation, and elicited imitation. The second part of the test is a 10-minute video interview, where test takers respond to four writing prompts and four speaking prompts. Test results include an overall score as well as subscores for Literacy, Conversation, Comprehension, and Production.

Duolingo English Test is an exam that has gained popularity in the last couple of years. Back in 2018, Plough et al. stated that tools such as Duolingo were starting to become more used, and in 2019 it was considered to be a low-stakes commercial language learning system (Zechner & Evanini, 2019). However, since the COVID-19 pandemic, the exam has gained much more popularity due to its "at home nature", and more universities have started accepting it as a measure of applicants' English language proficiency (Wagner, 2020). The exam has previously been criticized for presenting discrete items that are decontextualized in many cases (Wagner & Kunnan, 2015). Most of the research found about the test are technical reports that are listed in the Duolingo website, which serves as evidence that more research published in scientific journals is needed.

LanguageCert Test International ESOL

LanguageCert, a member of The PeopleCert Group that is regulated by Ofqual and Qualification Wales, provides different varieties of exams to accredit proficiency in English. The LanguageCert Test International ESOL is one of these options. There are different exams that test takers can choose from depending on the level (A1-C2), and for each level there are two exams: Written (Listening, Reading, Writing) and Spoken (Speaking). Levels A1 and A2 can only be taken at a testing center. The rest of the levels can also be taken online from the test taker's own computer with remote live proctoring. The price for the computer-based exams vary by level (B1: €50; B2: €80; C1: €90; C2: €100). However, the price of the paper-based version varies by testing center.

LanguageCert exams are being used for different purposes, including migration, work and study. They are recognized at the moment in 23 countries (mostly European). Some of the countries with the highest number

of institutions that accept LanguageCert Test International ESOL include Greece, Hungary, Italy, New Zealand and Spain. For each level's exam, the Listening, Reading and Speaking components have four different tasks, with 26 items for both Listening and Reading and four tasks for Speaking. The Written exam has two tasks. In order to pass, test takers must receive at least 75 out of 150 points on the Written exam and 25 out of 50 points on the Speaking part. Results are usually delivered within 3 business days for online exams, 5 business days for computer-based exams and 10 business days for paper-based exams. The length of the exam also varies by level, from about 1 hour and 50 minutes for level A1 to approximately 3 hours and 30 minutes for level C2. The duration of the exam is the same for all the versions (paper-based, computer-based or online). Computer based tests are recorded by the computer and sent as data asynchronously, whereas data is stored synchronously for online tests.

There is little research about this exam. In their review about LanguageCert, Isbell and Kremmel (2020) state that "as a relatively new suite of exams, there is limited evidence available pertaining to the relationship between exam content and academic English. Similarly, limited validation research is available at this time" (p. 7).

Linguaskill

Developed by Cambridge Assessment, Linguaskill is advertised on its website as a "quick and convenient online test to help organizations check the English levels of individuals and groups of candidates, powered by Artificial Intelligence technology." It is an online test that can be administered at the test taker's own venue with the help of a computer, internet connection, a microphone and a set of headphones. The exam, with a price that ranges from €70 to €90 (depending on the country and testing center), is used all over the world mostly by Higher Education Institutions to assess language levels on admissions, monitor progress, and check that students meet graduation language requirements. It is also used by employers to check applicants have the right language skills.

The test is divided into modules, and each module includes speaking, writing, reading and listening tasks. One of its main characteristics is that the Reading and Listening modules are adaptive, so there is not a certain number of questions that the test takers have to answer. The test finishes when enough questions have been answered for the test to identify the language proficiency level. The length for these two modules varies from 60 to 85 minutes. The Reading module includes questions such as read and select, gapped sentences, multiple-choice gap-fill, open gap-fill and extended reading. On the other hand, the Listening module includes listen and select, and extended listening questions. Results from these two modules are provided immediately.

The Writing module, which is scored automatically by the computer, lasts 45 minutes and includes two parts. In Part 1 (which lasts about 15 minutes), test takers read a short prompt and write an email of at least 50 words. Part 2 takes approximately 30 minutes and requires test takers to read a short text and write a response of at least 180 words. Answers are marked automatically by the computer, and results are available within 12 hours.

Finally, the Speaking component is done with the use of a microphone and headphones. The responses are recorded and then assessed by human examiners or market-leading auto-marking technology, known as hybrid marking. Results are available within 48 hours. This section lasts 15 minutes and has five parts, each part accounting for 20% of the final grade.
The number of studies involving Linguaskill is limited, and most of them are reports that have been conducted by Cambridge Assessment English. Back in the nineties, when Computer-Adaptive Tests started being developed, Linguaskill received much attention, as it was an innovative way of assessment that was created for Manpower Europe, a large employment services company. In Milanovic's words, the test "focus[ed] on language for work purposes but [was] also notable for the fact that it is a multi-lingual system operating in English, French, German, Spanish and Dutch and reporting on the same measurement scale"

(Chalhoub-Deville, 1999, p. VIII). Cheung et al. (2017) focus on the idea that it contains a writing assessment which is automatically marked by "a series of computer algorithms that has learned how to mark test responses from a large collection of learner responses marked by expert human markers" (p. 3). This idea is also supported by Seed and Xu (2017), who value the auto-marker's scoring accuracy, concluding that it is very satisfactory and reliable.

Oxford Test of English

Like Linguaskill, the Oxford Test of English is a computer adaptive proficiency test, with the difference that this test focuses only on the CEFR levels of A2, B1 and B2. Developed by Oxford University Press and certified by the University of Oxford, it covers the four skills (Speaking, Listening, Reading, and Writing). The Reading and Listening components are the adaptive modules, while the questions in the Speaking and Writing parts are randomized, so each test taker answers different questions. The test takes approximately 2 hours, and the scaled score ranges from 51 to 140. Results are usually given within 14 days of completion of the exam. The Oxford Test of English is recognized by universities, educational institutions and organizations around the world. The price varies depending on the number of modules completed, from €95 for one or two modules to €125 for three or four modules.

The Speaking section has four parts (interview, voicemails, talk and follow-up questions), with a total of 6 tasks and 15 items. This section lasts approximately 15 minutes. The Listening component lasts approximately 30 minutes and has four parts as well (multiple choice-picture options, note-completion, matching and multiple choice) with 12 tasks and 20 items. The Reading section has a duration of 35 minutes and includes four parts (multiple-choice questions on short texts, multiple matching, gapped sentences and multiple-choice questions on a longer text), with 9 tasks and 22 items. Finally, the Writing component has two parts (e-mail writing of approximately 80-130 words, and essay or article/review writing of approximately 100-160 words), with two tasks and two items.

No research has been found about this exam, and while the test specifications (OUP, 2020) describe the validation process with "over 10,000 students across thirty-seven countries from a wide number of first-language backgrounds at each of the targeted CEFR levels," studies are necessary to provide an objective view and to validate all the data and information provided by the test developer.

Pearson Test of English (PTE) Academic

The exam, developed by Pearson, is a computer-based exam designed to assess real-life, academic English, so passages and audio in the test are sourced from parts of lectures or any other academic materials. The range of accents included in the test also vary from American to British and non-native speakers. One of the characteristics of the exam is that it is an integrated skill test, so some of the tasks (while they are primarily assessing one particular skill) involve two skills, such as listening and writing or reading and writing. Due to its academic nature, scores are aligned not only with the CEFR levels A2-C2, but also the IELTS Academic test and TOEFL®. With an average length of about 3 hours and a price of $245 USD, test takers can do the exam at any Pearson test center, where they require a computer and a headset for the listening component. PTE Academic is accepted in many countries worldwide by thousands of universities. It is also a test used for visa purposes in countries such as Australia and New Zealand.

The test has three separate parts. Part 1, with eight different tasks, assesses Speaking and Writing (together). In Part 2, test takers have to complete five Reading tasks. Finally, Part 3 includes eight Listening tasks. The entire test has different types of items, such as multiple choice, fill in the blanks, re-ordering or short answers. Scores range between 10 and 90 points, and they must be interpreted carefully in terms of language proficiency. According to the PTE Score Guide for Test Takers, "if a test taker's PTE Academic score is 36, this predicts that they will perform successfully on the easiest tasks at B1. From 36 to 43, the likelihood of successfully performing the easiest tasks develops into doing well on the average tasks at B1" (Pearson, 2020, p. 28). Ranges between the easiest and most difficult task results at each level are provided to help

test takers determine their language proficiency level (see marks/points for the Pearson Test of English Academic in Table 3).

A review of the literature reveals several studies about the PTE Academic in the last few years (McCray & Brunfaut, 2016; Green, 2018; Barkaoui, 2019; Knoch et al., 2020; Rukthong & Brunfaut, 2020), which makes this test a source of research due to its applicability and impact in the academic field. In their review of the test, Wang et al. (2012) claim that the main use of admission to higher education has positive evidence, while they suggest some recommendations, such as "improving the quality of multiple-choice items or using a different test format [to] reduce the impact of test method on the intended score interpretations" (p. 617). It is important to note that, due to its "perceived importance and difficulty" (Knoch et al., 2020, p. 18), the test has some important consequences and therefore negative washback has been associated with it. Green (2018) focuses on aspects related to linking with the CEFR, concluding that "test score users should be clearly warned not to rely on CEFR level correspondences as a basis for high-stakes decision making" (p. 12).

TOEIC® Exams

The TOEIC® exam is developed by Educational Testing Service (ETS), the testing company that develops tests such as TOEFL® and PRAXIS. TOEIC assesses English language in the skills of Listening, Reading, Speaking and Writing, and they are assessed in two different tests: TOEIC® Listening and Reading, and TOEIC® Speaking and Writing. They are used in more than 160 countries and the price, approximately $85 USD for each component (Listening and Reading / Speaking and Writing), is set by the testing center. It is a popular exam in Asia, as evidenced by the research that has been conducted in countries such as Japan (In'nami & Koizumi, 2012, 2017), South Korea (Booth, 2017), or Vietnam (Nguyen & Gu, 2020).

The TOEIC® Listening and Reading test is a paper-and-pencil, multiple-choice assessment, and there are two timed sections of 100 questions each. The test takes approximately 2 hours and 30 minutes, with 45 minutes for Section I (Listening) and 75 minutes for Section II (Reading). In the Listening section, which has four different parts, test takers have to listen to a variety of questions and short conversations recorded in English, and then answer questions based on what they have heard (100 items total). In the Reading section, test takers have to read a variety of materials and respond to a total of 100 items, organized in three different parts. The TOEIC® Reading and Listening gives a score between 10 and 990. The TOEIC® Speaking and Writing test is an online test that is taken in a test center. This is a fairly new test which is only available in some countries. The TOEIC® Speaking test consists of 11 questions, as showing below, with a duration of 20 minutes. The score received in this part ranges from 0 to 200 points. Finally, the TOEIC® Writing test has eight questions, with a duration of approximately 60 minutes. The score also ranges from 0 to 200 points.

In their review about this exam, Im and Cheng (2019) state that its purpose "has been well achieved by using a very sophisticated method of domain analysis (i.e., the ECD approach) and by providing consistent test results across administrations of the TOEIC" (p. 322). Some studies focus on the correlation between the Reading and Listening factors of the test, which proves to be high (In'nami & Koizumi, 2012) or the prediction performance of test results on real-life English language tasks (Powers & Powers, 2015; Schmidgall & Powers, 2020). However, Im and Cheng (2019) claim the construct of the exam needs to be expanded in order to include more real-world language and tasks. This is also influenced by the use that is made of the TOEIC results. Like any other high-stakes tests, it can have a negative impact when results are used in certain ways (Booth 2017). The communicative skills the test is ostensibly testing for can be negated when test takers see the test only as an exit requirement for higher education. (Nguyen & Gu, 2020).

Trinity's Integrated Skills in English (ISE) Exams

The test is developed by Trinity College London and assesses the four skills (Reading, Writing, Speaking and Listening). The skills are divided in two exams: (1) Reading and Writing and (2) Speaking and

Listening, which can be completed together or separately. The levels are linked to the CEFR proficiency levels, and test takers can choose one of the following exams: ISE Foundation (A2), ISE I (B1), ISE II (B2), ISE III (C1) and ISE IV (C2).

One of the functions of this test is settlement and visa application for the UK. It is also used to be accepted in universities in the UK, Ireland and North America, and for end of study abroad programs. The price varies depending on the exam and level, ranging from €100 for ISE Foundation (both parts) to €210 for ISE IV.

The structure of the first four level exams is similar. The Reading & Writing exam, which lasts two hours, has four different tasks (long reading, multi-text reading, reading into writing and extended writing). The length of the Speaking & Listening exam varies depending on the level (from 13 minutes for ISE Foundation to 25 minutes for ISE III). This exam also has four different tasks (topic task, conversation task, and independent listening task 1 and 2). The structure of ISE IV, however, is different. It has three main components: a portfolio that allows approximately 6-12 weeks preparation time, a controlled written exam that lasts three hours and a 25-minute interview.

One of the concerns of Trinity College London has been the standard-setting process for the exam, and a couple of studies addressing this aspect have been carried out. The first one was a project conducted by Papageorgiou (2007), where he focused on factors and problems that judges consider when making decisions in the CEFR cut score setting process. His work has been published in different venues (Papageorgiou, 2010a, 2010b) and referenced by others (Taylor, 2009). A more recent study by Harsch and Paraskevi Kanistra (2020) presents a standard-setting approach to align some of the writing tasks to the CEFR. In their study, they conclude that their approach “enhances judgement validity and consequently alignment validity, as it allows panel facilitators to monitor how panelists use the CEFR descriptors and match them to task demands and performance features” (p. 19).

Table 1

Main Characteristics of APTIS and Duolingo English Test

	APTIS	**Duolingo English Test (DET)**
Developer	British Council	Duolingo, Inc.
Website	https://www.britishcouncil.es/en/exam/aptis	https://englishtest.duolingo.com
Delivery	Computer-based exam	Computer-adaptive exam
Length	Grammar and Vocabulary: 25 minutes Speaking: 12 minutes Writing: 50 minutes Listening: 40 minutes Reading: 35 minutes	Setup: 5 minutes Adaptive test (reading, writing, speaking and listening): 45 minutes Video interview: 10 minutes
Levels	CEFR levels: A1-C	CEFR levels: A1-C2
Parts/sections	Grammar and vocabulary: 60 tasks Speaking: 4 tasks Writing: 3 tasks Listening: 28 tasks (+/- 4) Reading: 4 tasks	There are five item types in the computer-adaptive portion of the test: c-test, audio yes/no vocabulary, visual yes/no vocabulary, dictation, and elicited imitation. Additionally, test takers respond to four writing prompts and four speaking prompts, which are not a part of the computer-adaptive portion of the test
Marks/points	An APTIS candidate will receive a score on a numerical scale (0-50) for the grammar and vocabulary section, and a score on a numerical scale (0-50) and CEFR level (A1 – C) for each skill they take and a CEFR (MERC) GLOBAL MARK	Scale from 10-160: 10-20: A1 25-55: A2 60-85: B1 90-115: B2 120-140: C1 145-160: C2 Test results include an overall score as well as subscores of Literacy, Conversation, Comprehension and Production
Results	Delivered within 48-72 hours No expiration date	Delivered within 48 hours Valid for two years

Table 2

Main Characteristics of LanguageCert Test International ESOL and Linguaskill

	LanguageCert Test International ESOL	**Linguaskill**
Developer	LanguageCert	Cambridge Assessment
Website	https://www.languagecert.org/welcome?gclid=EAIaIQobChMItt-U2vCn6wIVSflRCh10TQmBEAAYASAAEgLihfD_BwE	https://www.cambridgeenglish.org/exams-and-tests/linguaskill/
Delivery	Paper and pencil exam, computer-based exam or online exam	Online test that can be taken at own venue Reading and listening are adaptive tests Writing is scored automatically Speaking is recorded
Length	Listening: A1-A2: 20 minutes; B1-C2: 30 minutes Reading & Writing: A1-A2: 1 hour 20 minutes; B1-B2: 2 hours 10 minutes; C1-C2: 2 hours 40 minutes Speaking: A1: 6 minutes; A2: 9 minutes; B1: 12 minutes; B2: 13 minutes; C1: 15 minutes; C2: 17 minutes	Reading and Listening: 60-85 minutes Writing: 45 minutes Speaking: 15 minutes
Levels	CEFR levels: A1-C2	CEFR levels: A1-C2
Parts/sections	Listening and Reading: 4 parts with a total of 26 items in each part Speaking: 4 parts with 1 task in each part Writing: 2 parts with 1 task in each part	Reading: 5 types of questions and Listening has two types of questions Writing: 2 parts Speaking: 5 parts
Marks/points	Scale from 0-150 in the Written Exam and 0-50 in the Speaking part Candidates are awarded High Pass, Pass or Fail High Pass 101-150 / 150 35-50 / 50 Pass 75-100 / 150 25-34 / 50 Fail 0-74 / 150 0-24 / 50	Scale from 82-180: 82-99 (below A1) 100-11 (A1) 120-139 (A2) 140-159 (B1) 160-179 (B2) 180+ (C1 or above)
Results	Delivered within 3 business days for Online exams with remote, live proctoring 5 business days for computer-based exams 10 business days for paper-based exams No expiration date	Reading and Listening are delivered immediately Writing is scored automatically (results in 12 h) Speaking results within 48 hours Valid for two years

Table 3

Main Characteristics of Pearson Test of English (PTE) Academic and Oxford Test of English

	Pearson Test of English (PTE) Academic	**Oxford Test of English**
Developer	Pearson	Oxford University Press
Website	https://pearsonpte.com/pte-academic/	https://elt.oup.com/feature/global/oxford_test_of_english/?cc=global&selLanguage=en
Delivery	Computer-based exam	Computer-adaptive test
Length	Speaking & Writing: 77-93 minutes Reading: 32-40 minutes Listening: 45-57 minutes	Speaking: Approximately 15 minutes Listening: Approximately 30 minutes Reading: 35 minutes Writing: 45 minutes
Levels	CEFR levels: A2-C2	CEFR levels: A2, B1 and B2
Parts/sections	Three separate parts: Part 1. Speaking & Writing: 38-47 items Part 2. Reading: 15-20 items Part 3. Listening: 17-25 items	Speaking: 4 parts with a total of 6 tasks and 15 items Listening: 4 parts with 12 tasks and 20 items Reading: 4 parts with 9 tasks and 22 items Writing: 2 parts with 2 tasks and 2 items
Marks/points	Score range between 10-90 points. The ranges below present the easiest and most difficult tasks at each level: C2: 80-90 (average 85) C1: 67-84 (average 76) B2: 51-75 (average 59) B1: 36-58 (average 43) A2: 24-42 (average 30)	Scale from 51-140 points 51-81 (A2) 82-111 (B1) 112-140 (B2)
Results	Delivered within 48 hours Valid for two years	Delivered within 14 days Results remain valid and available until a new Oxford Test of English is taken

Table 4

Main Characteristics of TOEIC® Exams and Trinity's Integrated Skills in English (ISE) exams

	TOEIC® Exams	**Trinity's Integrated Skills in English (ISE) exams**
Developer	Educational Testing Service (ETS)	Trinity College London
Website	https://www.ets.org/toeic/	https://www.trinitycollege.com/qualifications/english-language/ISE
Delivery	TOEIC® Listening and Reading test is a paper-and-pencil test to be done in a testing center TOEIC® Speaking and Writing is an online test to be done in a testing center	Paper and pencil exam taken under exam conditions at Trinity registered centers
Length	TOEIC® Listening and Reading test: 2 hours and 30 minutes TOEIC® Speaking and Writing: 20 minutes for speaking and 60 minutes for writing	Reading and Writing: 2 hours Speaking and Listening: 20 minutes
Levels	Speaking: 8 proficiency levels Writing: 9 proficiency levels Listening & Reading: Scores are determined by the number of correct answers, which is converted to a scaled score	CEFR levels: A2-C2
Parts/sections	Two separate tests: TOEIC® Listening and Reading, with 4 sections for listening and 3 sections for reading TOEIC® Speaking and Writing, with 11 questions for speaking and 8 questions for writing	Reading and Writing test has 4 tasks Speaking and Listening test has 4 tasks
Marks/points	TOEIC® Listening and Reading test: 10-990 points in total TOEIC® Speaking and Writing: 0-200 points for each part	Maximum scores: Reading: 30 points Writing: 28 points Speaking: 19 points Listening: Points vary in each level (from 4 to 10 points)
Results	Delivered within 10 days Valid for two years	Candidates need to achieve at least the Pass score in each of the relevant skill areas to be awarded a module certificate The level of achievement (Distinction, Merit, Pass or Fail) for each of the four skills are stated on the qualification certificate, but these are not conflated to give an overall level of achievement Delivered within 21 days No expiration date

Validity of Multilevel Tests as an Enhanced Opportunity

There is no question that multilevel tests offer a large number of benefits for medium-stakes language testing. Innovation, however, may not always be as beneficial for the validity of these tests. In the report of the APTIS test by García Laborda et al. (2017), the authors mention that there are many reactions to speaking prompts that may jeopardize the candidate's performance and, thus, introduce issues that affect the consequential validity of the test. Therefore, it is necessary to revise the validity of these multilingual tests and their value in the context of an argument-based approach.

As we mentioned before, multilingual tests usually use pre-validated items either from other exams or from the test publishers. Although there have been a number of claims that suggest that computer-based language testing should be looking at a new construct definition of what knowing a language means in the 21st century, the new complex systems of scoring and rating make possible the inclusion of new types of tasks such as cooperative and online tasks. Multilevel tests have, however, permitted the implementation of traditional items in a way that is faster and more efficient. Nevertheless, what is still missing are the necessary algorithms that promote learning and not just diagnostic exercises. This very much means that these tests should adapt to personalize the report for each specific test taker.

Validity needs to be the fundamental concern looked for when innovating testing systems, and the fundamental warrant of their own use for the purpose of measuring the candidate's academic competence in a foreign or second language. As we mentioned, validation must be organized in relation to Kane's (2009) argument-based approach (also Weir, 2005). According to this, multilevel tests already consider the consequential validity based on the assumptions of the interpretation and use of the score or competence level they assign. This provides them with a significant information transfer towards society. This means, in simple language, that the tests we suggested in this paper have been proven to obtain evidences that support the candidate's results. While it is true that some of these may have a "better reputation" than others based on the item supply, question randomization, report issuer selection and more, all of these have been conveniently accredited and are widely accepted in countries and institutions worldwide. However, what is still missing in many of them is a real and sound corpus of research especially in aspects such as external validity, generalization, extrapolation and decision. Research on the topic could actually lead to dramatic changes and the revision of some of the certification tests (mainly IELTS and Ib TOEFL).

The Future of Multilevel Tests

As mentioned above, the authors of this paper consider that the types of items that these multilevel tests deal with are based on an old constructivist model that has been revised and improved for a number of years but has ignored the evolution of language learning, especially through technology. It is really hard to understand today's language as an isolated knowledge rather than as the cooperation of foreign and native speakers, the interaction with the Internet and its supporting tools, cooperation in writing design and implementation (especially of documents), the interaction with specific fields of study (Content Language Integrated Learning, or CLIL), the use of language for reasoning and many other issues that also limit the application of the consequential validity of "knowing a language". Innovation must be seen in light of, at least, three categories: (a) items or tasks; (b) test construction, assembly and delivery; and (c) innovations and personal factors. At least in educational contexts, language tests should consider measuring competence in these 21st century skills (although they may evolve in the light of the 2020 use of technology due to the COVID-19 pandemic). In relation to new types of assessments, body language must be also measured in online speaking assessments.

Looking at specific current deficits in item design, while it is true that body language varies a great deal among users, in more than a few cases (especially with beginner students) it has a significant role in communication. Furthermore, it also enhances online synchronic communication. Publishers should be

looking at new types of items. For example, the integrated approach used since 2007 by Ib TOEFL led to new ways to construct assessment for the prospective capacity of a student in an academic environment. No matter whether a test looks at academics or general use of the language, new language tasks to prove the students' competence as well as their capacity to use a different language also need to be measured. This could be improved by more use of simulations (instead of just delivering a video), cooperative problem solving or mini presentations coordinated online, and the use of online reference materials (similar to Wikipedia or *ad hoc* documents). All these types of items go beyond the traditional wrong/right or even assessment of a programmed pair conversation. All of these can be considered as hybrid items since they require the integration of more than just one skill.

In relation to test construction, these multilevel tests use Computer-Adaptive systems. Usually, the problem is that the randomization of items may cause problems since the same one item may be brought to the test takers often, creating the feeling that the same item is used just too commonly and introducing a risk to the transference of the same item. Therefore, an adequate pool size may not be enough. An automated test generator may help but only just to structure the test, not to increase its validity. About the delivery, although much has been done in relation to online proctoring, there is still some hard work to be done to respect the different privacy rights which test users may have in different countries. For instance, during the pandemic in Spain a student in a public university presented charges against its university because they wanted to access his home remotely for a test.

Finally, in relation to the learning opportunities, it is undeniable that a test, no matter its nature, should serve to identify real learning needs. These multilevel tests do not help much to orientate further learning. Therefore, the reports should aim not only to just give a final score or summary of competence indicators but also to providing more information on the specifics that need to be either revised or improved. In this sense, learning analytics can be used not only to reinforce but also to give a social application for creating error banks or collections as well as creating patterns of learning across countries and different groupings of people (García Laborda, 2017).

Additionally, external validation studies are missing. Most of the information that is received by the different stake holders actually comes from the experience or comparability between tests from the same publisher, say ETS or Cambridge assessments just to mention a few. However, no external validation studies have been done. We also mentioned item sampling as another issue at stake. The consequential validity (also known as extrapolation inference) needs comparisons with real life tasks but this is an aspect that has been commonly neglected in language testing.

In relation to the tests' validity, publishers acknowledge that they are based on the use of items used for the certification tests apart from internal validity (across the different skills plus grammar and vocabulary). In reality, some of these tests are "informally" considered "softer" than others while they would have the same value of external impact in universities and educational and professional boards of different stake holders (consequential validity). However, sound studies and further research are necessary in order to consider seriously these popular beliefs.

Conclusion

In this paper, we stated the values, definition, construction and features of some (there are many more) international multilevel tests used for accreditation of competence levels in English. We also addressed several technological issues involved in the development of those tests. As we have seen, the publishers' experience seems to be the most important warrant of their own quality. However, although more research on these tests is absolutely necessary, very little attention has been given to them despite their importance (especially in educational settings). Apart from external and consequential validity, one can think that scoring and generalization are the current major concerns when addressing these tests overall.

It has also been suggested that computers should lead to a significant revolution in item and construct design since most of the items lead to an individual knowledge of the language when, in fact, communication is co-constructed and, in such a sense, the traditional 5-minute artificial speech of pair-dialogues found in many tests or delivering a monologue to a "machine" may not suffice.

Furthermore, since the construct of each test is different due to the changes in the inferences, claims and assumptions relevant to its use as well as its intended interpretation and use of test scores, it would probably be desirable not to mix them. We still remember a Master's program that used BULATS (a business test) as an access test for a Master of Education Degree in a very reputed university in Spain. Obviously, using the wrong tool might have led to wrong access decisions.

One other innovation that is missing is the real analysis of testing engines, which has little if any presence in published journals. This itself really weakens the professionals' belief that internal processes are clear, relevant and adequate for the test purpose. In this context, it is necessary to use and being able to replicate evidence-based techniques proven to reinforce the claims of each test. However, one of the major problems is that, in general, those who take decisions on the validity of certain tests as evidence of usefulness for social and academic purposes have neither the knowledge nor the skills necessary to have a critically founded opinion. We strongly believe that the future of language testing relies on technology but without more research, we must believe in what our eyes can't see.

Acknowledgements

The authors would like to express their gratitude to Mrs. Slavka Madarova for her revision and comments on this paper.

References

Bardovi-Harlig, K., & Shin, S. Y. (2014). Expanding traditional testing measures with tasks from L2 pragmatics research. *Iranian Journal of Language Testing*, *4*, 26–49.

Barkaoui, K. (2019). Examining sources of variability in repeaters' L2 writing scores: The case of the PTE Academic writing section. *Language Testing*, *36*(1), 3–25.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*, 355–377.

Booth, D. K. (2017). *The Sociocultural Activity of High Stakes Standardised Language Testing: TOEIC Washback in a South Korean Context*. Springer International.

British Council. (2021). *Assessment Research Grants.* https://www.britishcouncil.org/exam/aptis/research/assessment-advisory-board/awards/assessent-grants

Carr, N. T., & Xi, X. (2010). Automated scoring of short-answer reading items: Implications for constructs. *Language Assessment Quarterly*, *7*, 205–218.

Chalhoub-Deville, M. (Ed.). (1999). *Issues in computer-adaptive testing of reading proficiency*. Cambridge University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology.* Cambridge University Press.

Cheung, K., Xu, J., & Lim, G. (2017). Linguaskill: Writing trial report. *Cambridge English.* https://www.eiken.or.jp/linguaskill/assets/pdf/linguaskill-writing-trial-report.pdf

Duolingo, Inc. (2020, June 24). *Duolingo English Test: Security, Proctoring, and Accommodations*. https://duolingo-papers.s3.amazonaws.com/other/det-security-proctoring-whitepaper.pdf

Elder, C., & Randow, J. V. (2008). Exploring the utility of a web-based English language screening tool. *Language Assessment Quarterly*, *5*, 173–194.

Fernández Álvarez, M. (2016). Language testing in the digital era. In E. Martín-Monje, I. Elorza, & B. García Riaza (Eds.), *Technology-enhanced language learning for specialized domains. Practical applications and mobility* (pp. 61–72). Routledge.

García Laborda, J. (2007). On the net: Introducing standardized EFL/ ESL exams. *Language Learning and Technology*, *11*(2), 3–9. https://www.lltjournal.org/item/2567

García Laborda, J., Litzler, M. F., García Esteban, S., Pizarro, M. A., & Otero de Juan, N. (2017). Student perceptions of the CEFR levels and the impact of guided practice on Aptis Oral test performance. *ARAGs Research Reports Online*, APTIS AR-G/2017(5) https://www.britishcouncil.org/sites/default/files/laborda_et_al_layout_0.pdf

Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly*, *15*(1), 59–74.

Harsch, C., & Paraskevi Kanistra, V. (2020). Using an innovative standard-setting approach to align integrated and independent writing tasks to the CEFR. *Language Assessment Quarterly*, *17*(3), 262–281.

Holzknecht, F., McCray, G., Eberharter, K., Kremmel, B., Zehentner, M., Spiby, R., & Dunlea, J. (2020). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing*, *0*(0), 1–21.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practices*, *20*, 16–25.

Im, G., & Cheng, L. (2019). The Test of English for International Communication (TOEIC®). *Language Testing*, *36*(2), 315–324.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing*, *29*(1), 131–152.

In'nami, Y. & Koizumi, R. (2017). Using EIKEN, TOEFL, and TOEIC to Award EFL Course Credits in Japanese Universities. *Language Assessment Quarterly*, *14*(3), 274–293.

Isbell, D. R., & Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *0*(0), 1–21.

Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 39–64). Information Age Publishing.

Knoch, U., Fairbairn, J., & Huisman, A. (2016). An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test. *Papers in Language Testing and Assessment*, *5*(1), 90–106.

Knoch, U., Huisman, A., Elder, C., Kong, X., & McKenna, A. (2020). Drawing on repeat test takers to study test preparation practices and their links to score gains. *Language Testing*, *0*(0), 1–23.

Long, A. Y., Shin, S. Y., Geeslin, K., & Willis, E. W. (2018). Does the test work? Evaluating a web-based language placement test. *Language Learning & Technology*, *22*(1), 137–156. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44585/1/22_01_longshingeeslinwillis.pdf

Lowman, R. (2017). Ethical and legal concerns in internet-based testing. In J. Scott, D. Bartram, & D. Reynolds (Eds.), *Next generation technology-enhanced assessment: Global perspectives on*

occupational and workplace testing (Educational and psychological testing in a global context) (pp. 350–374). Cambridge University Press.

McCray, G., & Brunfaut, T. (2016). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing*, *35*(1), 51–73.

Nguyen, H., & Gu, Y. (2020). Impact of TOEIC listening and reading as a university exit test in Vietnam. *Language Assessment Quarterly*, *17*(2), 147–167.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*, 517–537.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *Modern Language Journal*, *93*, 836–847.

OUP (2020). Oxford Test of English test specifications. Oxford University Press.

Papageorgiou, S. (2007). *Relating the Trinity College London GESE and ISE exams to the Common European Framework of Reference: Piloting of the Council of Europe draft Manual Final Project Report.* Trinity College London.

Papageorgiou, S. (2010a). Investigating the decision-making process of standard setting participants. *Language Testing*, *27*(2), 261–282.

Papageorgiou, S. (2010b). Linking international examinations to the CEFR: The Trinity College London experience. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 145–157). Cambridge University Press.

Pearson (2020). *Score guide for test takers. Version 12 (April 2020)*. https://pearsonpte.com/wp-content/uploads/2020/04/Score-Guide-16.04.20-for-test-takers.pdf

Plough, I., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing*, *35*(3), 427–445.

Powers, D. E., & Powers, A. (2015). The incremental contribution of TOEIC® Listening, Reading, Speaking, and Writing tests to predicting performance on real-life English language tasks. *Language Testing*, *32*(2), 151–167.

Roever, C. (2001). Web-based language testing. *Language Learning & Technology*, *5*(2), 84-94. https://scholarspace.manoa.hawaii.edu/bitstream/10125/25129/1/05_02_roever.pdf

Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing*, *37*(1), 31–53.

Schmidgall, J., & Powers, D. E. (2020). Predicting communicative effectiveness in the international workplace: Support for TOEIC® Speaking test scores from linguistic laypersons. *Language Testing*, *0*(0), 1–24.

Seed, G., & Xu, J. (2017). Integrating technology with language assessment: Automated speaking assessment. In E. Gutiérrez Eugenio (Ed.), *ALTE, Learning and assessment: Making the connections – Proceedings of the ALTE 6th International Conference* (pp. 286–291). Association of Language Testers in Europe.

Shin, S. Y. (2012). Web-based language testing. In C. Coombe, B. O'Sullivan, P. Davidson, & S. Stoynoff (Eds.), *The Cambridge guide to language assessment* (pp. 274–279). Cambridge University Press.

Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2017). Scoring validity of the Aptis Speaking test: Investigating fluency across tasks and levels of proficiency. *ARAGs Research Reports Online*, *AR-G/2017*(7).

Taylor, C. (2009) Relating the Trinity College London International ESOL examinations to the CEFR. In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives* (pp. 69–74). Cito, Institute for Educational Measurement.

Valencia Robles, J. (2017). Anxiety in language testing: The APTIS case. *Profile: Issues in Teachers' Professional Development, 19*(1), 39–50.

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*(4), 493–513.

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

Wagner, E., & Kunnan, A. J. (2015). The Duolingo English Test [Test Review]. *Language Assessment Quarterly*, *12*(3), 320–331.

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson Test of English Academic: Building an assessment use argument. *Language Testing*, *29*(4), 603–619.

Weir, C. J. (2005). *Language Testing and Validation: An evidence-based approach*. Palgrave-Macmillan.

Williamson, D. M., Bejar, I. I., & Sax, A. (2004). Automated tools for subject matter expert evaluation of automated scoring. *Applied Measurement in Education*, *17*, 323–357.

Zechner, K., & Evanini, K. (Eds.). (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.

Appendix A. Structure of the Exams

Table 1
Comparison of APTIS and DET

	APTIS	Duolingo English Test (DET)
Core Part	Part 1. Grammar. Sentence completion (25 question with 3-option multiple choice) Part 2. Vocabulary. Sets of 5 target words with 10 options (25 questions in total) (1) Word matching (similar meaning). Match words to definitions. (2) Sentence completion. (3) Word pairs or word combinations (words commonly used together).	Tasks in the adaptive test assess integrated skills (literacy -reading & writing-, comprehension -reading & listening-, conversation -listening & speaking- and production -writing & speaking-) with different item types: • C-test: The first and last sentences are fully intact, while words in the intervening sentences are "damaged" by deleting the second half of the word. Test takers respond to the c-test items by completing the damaged words in the paragraph • Yes/no vocabulary: Test takers are presented with a set of English words mixed with pseudowords that are designed to appear English-like, and must discriminate between them • Dictation: Test takers listen to a spoken sentence or short passage and then transcribe it • Elicited imitation -read aloud-: The read-aloud variation of the elicited imitation task is a measure of test taker reading and speaking abilities. It
Listening	Part 1. Information recognition. Listen to a short monologue or dialogue to identify specific information (3-option multiple choice) Part 2. Information matching. Match people's monologues to information (6 pieces of information with 4 monologues) Parts 3+4. Inference. Listen to monologues and dialogues and identify the attitude, opinion or intention (3-option multiple choice)	
Reading	Part 1. Sentence comprehension. Choose words to complete sentences (5 sentences with 3-option multiple choice items) Part 2. Text cohesion. Put sentences into the correct order (2 tasks: 6 sentences jumbled up in each task Part 3. Opinion matching. Match people's opinions	

	to statements (7 statements matched to 4 people's opinions) Part 4. Long text comprehension. Match headings to paragraphs (8 paragraphs and 7 headings)	requires the test takers to read, understand, and speak a sentence • Extended speaking: At the end of the CAT portion of the test, the test takers respond to four speaking prompts: one picture description task and three independent speaking tasks, two with a written prompt and one with an aural prompt • Extended writing: Test takers respond to four writing prompts that require extended responses: three picture description tasks and one independent task with a written prompt
Writing	Part 1. Word-level writing. Respond to messages using individual words (1–5 words for each question) Part 2. Short text writing. Write personal information (20–30 words) Part 3. Three written responses to questions. Respond to written questions on a social network-type website (30–40 words for each question) Part 4. Formal and informal writing. Write an informal email to a friend and a formal email to an unknown person (40–50 words for the informal email and 120–150 words for the formal email)	
Speaking	Part 1. Personal information. Respond to three personal information questions (30 seconds for each response) Part 2. Describe, express opinion and provide reasons and explanations. Describe a picture and answer two additional questions of increasing difficulty (45 seconds for each response) Part 3. Describe, compare and provide reasons and explanations. Describe two contrasting pictures and answer two additional questions of increasing difficulty (45 seconds for each response) Part 4. Discuss personal experience and opinion on an abstract topic. Answer three questions on an abstract topic (1 minute to prepare and 2 minutes response time)	

Table 2

Comparison of LanguageCert International ESOL and Linguaskill

	LanguageCert International ESOL	**Linguaskill**
Listening	Part 1. Recognize simple key information in short statements (A1-A2). Understand context, meaning and function of a range of utterances (B1) or in short conversations on concrete and abstract topics (B2-C2) Part 2. Identify functions in short utterances typical of spoken English (A1-A2). Identify a specific aspect of a conversation (B1-C2) Part 3. Identify a specific aspect of a conversation (A1-A2). Extract key information from a monologue to complete a task (B1-C2) Part 4. Extract key information from a dialogue (A1) or monologue (A2). Follow a discussion between two speakers (B1-C2)	Listen and select. Candidates listen to a short audio recording and answer a multiple-choice question with three options Extended listening. Candidates listen to a longer recording and answer a series of multiple-choice questions based on it. The questions are in the same order as the information they hear in the recording
Reading	Part 1. Understand the organizational and lexical features of the text (A1). Understand coherence and cohesion of short texts (A2) and a variety of authentic texts (B1). Understand in detail information, ideas and opinions (B2). Understand articles, use of language and texts dense with complex structures (C1). Understand literary texts,	Read and select. Candidates read a notice, label, memo or letter containing a short text and choose the sentence or phrase that most closely matches the meaning of the text. There are three possible answers Gapped sentences. Candidates read a sentence with a missing word (gap) and choose the correct word to fill the gap. There are three or four choices for each

	use of emotive language and texts dense with complex structures (C2) Part 2. Understand the structure of a short simple text (A1-A2). Understand how meaning is built up in a text (B1-C2) Part 3. Understand the purpose of text and to locate specific information (A1-B1) and awareness of writers' stance and attitude (B2-C2) Part 4. Identify meaning in short texts (8 short texts) (A1). Understand specific information through detailed reading (A2-B2). Understand text discourse, purpose and gist and to locate specific information (C1-C2)	gap Multiple-choice gap-fill. Candidates choose the right word or phrase to fill the gaps in a text. There are three or four choices for each gap Open gap-fill. Candidates read a short text in which there are some missing words (gaps) and write in the missing word in each gap Extended reading. Candidates read a longer text and answer a series of multiple-choice questions. The questions are in the same order as the information in the text
Writing	Part 1. Communicate ideas or basic information (A1: 4 sentences in about 30 words; A2: 30-50 words). Respond appropriately to a given text in order to produce a formal response for an intended public audience (B1: 70-100 words; B2: 100-150 words; C1-C2: 150-200 words) Part 2. Produce short simple text for an intended audience (A1: 20-30 words; A2: 30-50 words). Produce an informal letter to a friend (B1: 100-120 words). Produce a personal letter, a narrative composition/ story or a descriptive composition (B2: 150-200 words; C1-C2: 250-300 words)	Part 1. Candidates read a short prompt, usually an email. They use the information in the prompt and the three bullet points to write an email of at least 50 words. Part 2. Candidates read a short text outlining a scenario and respond using the information in the scenario and the three bullet points. Candidates will write at least 180 words to a wider audience and may be asked to produce a variety of text types (e.g. review, article, web post)
Speaking	Part 1. Give and spell name. Give country of origin. Answer five questions Part 2. Two or three situations are presented by the interlocutor at each level and candidates are required to respond to and initiate interactions Part 3. Exchange information to identify similarities and differences in pictures of familiar situations at Preliminary and Access levels. Hold a short discussion to make a plan, arrange or decide something using visual prompts at Achiever, and written text as the prompt at the three higher levels Part 4. After 30 seconds of preparation time, talk about a topic provided by the interlocutor and answer follow-up questions (A1: half a minute; A2: 1 minute; B1: 1 and a half minutes; B2: 2 minutes; C1: 2 minutes; C2: 3 minutes)	Part 1. Interview (8 questions). The candidate answers eight questions about themselves (the first two questions are not marked) Part 2. Reading aloud (8 questions). The candidate reads eight sentences aloud Part 3. Long turn 1 (1 question). The candidate is given a topic to talk about for 1 minute Part 4. Long turn 2 (1 question). The candidate is given one or more graphics (for example a chart, diagram or information sheet) to talk about for 1 minute Part 5. Communication activity (5 questions). The candidate gives their opinions in the form of short responses to five questions related to one topic

Table 3

Comparison of Pearson Test of English (PTE) Academic and Oxford Test of English

	Pearson Test of English (PTE) Academic	**Oxford Test of English**
Listening	Part 1. Summarize spoken text (listening & writing). Test takers hear an audio recording and need to write a 50-70-word summary on what they heard (10 minutes) Part 2. Multiple choice choose multiple answer (listening). Test takers need to listen to a recording and answer multiple-choice questions. There is more than one correct response Part 3. Fill in the blanks (listening & writing). Test takers are presented with a transcript of an audio recording, but some words are missing. They have to	Part 1. Multiple choice – picture options. Five short monologues/ dialogues each with one 3-option multiple-choice question with picture options Part 2. Note-completion. A longer monologue with five 3-option multiple-choice note-completion questions Part 3. Matching opinions with people who say them. A longer dialogue with five 3-option multiple-choice questions focusing on identifying opinion Part 4. Multiple choice. Five short monologues/

	restore the transcript by typing in the missing words Part 4. Highlight correct summary (listening & reading). Test takers need to select the summary that best matches a recording Part 5. Multiple choice, single answer (listening). Test takers need to listen to a recording and answer multiple-choice questions Part 6. Select missing word (listening). The last word or group of words in a recording has been replaced by a beep. Test takers need to select the most appropriate option to complete the recording. Part 7. Highlight incorrect words (listening & reading). Test takers are presented with the transcript of an audio recording, but the transcript contains some errors. While listening and reading, test takers need to select the words in the text that differ from what the speaker says Part 8. Write from dictation (listening & writing). Test takers hear a short sentence. They need to type the sentence into the response box at the bottom of the screen	dialogues each with one 3-option multiple-choice question Abilities assessed: identifying main meaning; identifying details; global and local meaning; identifying opinion and attitude; understanding implied meaning, interaction, and pragmatics
Reading	Part 1. Reading & writing (reading). Fill in the blanks. Test takers need to select the most appropriate words from a drop-down list to restore a text Part 2. Multiple choice, multiple answers (reading). Test takers need to read a passage and answer multiple-choice questions. There is more than one correct response Part 3. Re-order paragraphs (reading). Test takers need to restore the original order of a text by selecting text boxes and dragging them across the screen Part 4. Fill in the blanks (reading). Test takers need to drag and drop words across the screen to correctly fill in the gaps in a text Part 5. Multiple choice, single answer (reading). Test takers need to read a passage and answer multiple-choice questions	Part 1. Multiple-choice questions on short texts. Six short texts from a variety of sources, each with one 3-option multiple-choice question Part 2. Multiple matching. Six profiles of people to match with four longer text descriptions Part 3. Gapped sentences. Six extracted sentences are inserted into a longer text Part 4. Multiple-choice questions on a longer text. Four 3-option multiple-choice questions Abilities assessed: careful reading; expeditious search reading; local and global meaning Inference; understanding attitude, opinion, and writer purpose; understanding reference and meaning in context
Writing	Part 1. Summarize written text (reading & writing). Test takers need to write a summary of a text in one sentence. They have 10 minutes to write their summary, in which they have to include the main points of the reading passage in a full, single sentence of no more than 75 words Part 2. Essay (writing). Test takers need to write a 200-300 word argumentative essay in response to a prompt. They have 20 minutes to write the essay	Part 1. Email. Written response to an input email; 80 – 130 words Part 2. Essay or article/review. Essay OR article/review on a topic typical of classroom discussion; 100 – 160 words Abilities assessed: giving information; expressing and responding to opinions and feelings; inviting, requesting, and suggesting; writing to develop and argument; narrating and describing; writing to persuade or suggest
Speaking	Part 1. Personal introduction (speaking). Test takers need to give some personal information for 30 seconds. This item is not scored Part 2. Read aloud (reading & speaking). Test takers need to read a written text aloud Part 3. Repeat sentence (listening & speaking). Test takers need to repeat the sentence they hear Part 4. Describe image (speaking). Test takers need to describe an image. They have 25 seconds to study	Part 1. Interview. Eight questions on everyday topics Part 2. Voicemails. Two voicemails in response to two different situations Part 3. Talk. Short talk on an issue or scenario Part 4. Follow-up questions. Six follow-up questions on the theme of the Part 3 talk Abilities assessed: responding appropriately to

	the image and prepare the response	questions; giving factual information
	Part 5. Re-tell lecture (listening & speaking). Test takers need to re-tell what they heard. They may also see an image related to the audio	organizing extended discourse; describing, comparing, contrasting, speculating, and suggesting
	Part 6. Answer short question (listening & speaking). Test takers need to reply to a question in one or a few words. They may also see an image related to the audio	

Table 4
Comparison of TOEIC® Exams and Trinity's Integrated Skills in English (ISE) exams

	TOEIC® Exams	Trinity's Integrated Skills in English (ISE) exams
Listening	Al parts: Understand spoken English Part 1. Photographs (6 questions). Select the statement that best describes what is shown in a picture Part 2. Question-Response (25 questions). Select a response to a question or statement which are not printed Part 3. Conversations (39 questions; 13 conversations with 3 questions each). Answer 3 questions about what a speaker says in each conversation Part 4. Short Talks (30 questions; 10 talks with 3 questions each). Answer 3 questions about what a speaker says in each talk	Task 1. Independent listening. The examiner introduces the talk and then the recording will play once. After the first time the test taker tells the examiner in one or two sentences what the talk is about. The examiner will then ask the test taker a question about the talk and provide some paper. The test taker listens again and takes some notes. The examiner will ask the question again and the test taker responds to the examiner's question for up to one minute
Reading	Part 5. Incomplete Sentences (30 questions). Select the best answer from 4 choices to complete a sentence with a word or phrase missing Part 6. Error Recognition or Text Completion (16 questions). Select the best answer from 4 choices to complete some incomplete sentence in a text with a word or phrase missing Part 7. Reading Comprehension. Answer comprehension questions after reading a selection of texts, such as magazine and newspaper articles, e-mails, and instant messages. Single passages: 29 questions; 10 reading texts with 2-4 questions each Multiple passages: 25 questions; 5 sets of double or triple passages with 5 questions per set	Task 1. Long reading. 15 questions: Title matching: For questions 1–5 test takers must choose a title for each paragraph Selecting true statements: For questions 6–10 test takers must decide which five statements from a list of eight are true Completing sentences: For questions 11–15 test takers must choose an exact number, word or phrase (maximum three words) from the text to complete gaps Task 2. Multi-text reading. 15 questions: Multiple matching: For questions 16–20 test takers must read four texts and think how they would summarize each text. They have to read the questions — each question refers to one of the four texts. Choose which text matches the questions Selecting the true statements: For questions 21–25 test takers must decide which five statements from a list of eight are true Completing the notes section: For questions 26–30 test takers must choose an exact number, word or phrase (maximum three words) from the text to complete gaps Task 3. Reading into writing: Test takers must read four texts and use information from the texts from task 2 to write an answer to a question
Writing	Questions 1-5. Write a sentence based on a picture. Write 1 sentence based on a picture. With each picture, 2 words or phrases that must be used are given (Grammar, relevance of the sentences to the pictures) Questions 6-7. Respond to a written request. Show how well test takers can write a response to an email (Quality and variety of sentences, vocabulary, organization) Question 8. Write an opinion essay. Write an essay in response to a question that asks to state, explain and support the opinion on an issue (Whether	Task 1. Extended writing. Write a short text similar to the kind of writing done in school or college

	opinion is supported with reasons and/or examples, grammar, vocabulary, organization)	
Speaking	Questions 1-2. Read out loud a text on the screen (Pronunciation, intonation and stress) Question 3: Describe a picture on the screen in as much detail as possible (All of the above + grammar, vocabulary and cohesion) Questions 4-6. Answer 3 questions (All of the above + relevance of content and completion of content) Questions 7-9. Answer 3 questions based on information provided (All of the above) Question 10. Propose a solution to a problem that is presented (All of the above) Question 11. Give the opinion about a specific topic (All of the above)	Task 1. Topic. Test takers must choose a topic they are interested in (anything they can talk about) and prepare by writing a mind map and think of different areas to talk about related to the topic for up to 4 minutes Task 2. Collaborative. The examiner reads a prompt and the test taker needs to ask questions and make comments to keep the conversation going. They need to keep the conversation going in this task, think about ways to ask questions, get more information and clarify details for up to 4 minutes Task 3. Conversation. The examiner chooses one subject areas from a list, and will ask the test taker about the subject. The conversation will last up to 2 minutes

About the Authors

Jesús García Laborda has a Doctorate in English Philology and a European Doctorate in Didactics. He is currently the Dean of the College of Education of Universidad de Alcalá and the research head of Bilingual Education at Instituto Franklin (UAH). He has published over 200 papers in CALL, CALT, ESP, bilingual education and teacher education.

E-mail: jesus.garcialaborda@uah.es

Miguel Fernández Álvarez is an Assistant Professor in the Department of Linguistics Applied to Science and Technology at the Universidad Politécnica de Madrid (UPM), where he teaches English for Specific Purposes within the field of construction. His research areas include bilingual education, second language acquisition and language assessment.

E-mail: m.fernandez@upm.es