International Journal of Listening

ISSN: 1090-4018 (Print) 1932-586X (Online) Journal homepage: https://www.tandfonline.com/loi/hijl20

Reviews of Academic English Listening Tests for Non-Native Speakers

Tingting Kang, Maria Nelly Gutierrez Arvizu, Panjanit Chaipuapae & Roman Olegovich Lesnov

To cite this article: Tingting Kang, Maria Nelly Gutierrez Arvizu, Panjanit Chaipuapae & Roman Olegovich Lesnov (2019) Reviews of Academic English Listening Tests for Non-Native Speakers, International Journal of Listening, 33:1, 1-38, DOI: 10.1080/10904018.2016.1185210

To link to this article: https://doi.org/10.1080/10904018.2016.1185210

Published online: 27 Jun 2016.

International Journal of Listening, 33: 1–38, 2019

ISSN: 1090-4018 print / 1932-586X online
DOI: 10.1080/10904018.2016.1185210

Routledge
Taylor & Francis Group

Reviews of Academic English Listening Tests for Non-Native Speakers

Tingting Kang, Maria Nelly Gutierrez Arvizu, Panjanit Chaipuapae, and Roman Olegovich Lesnov

Northern Arizona University

This article presents a review of 20 tests designed for assessing the academic English listening skill of second or foreign language learners. The available test information has been systematically condensed in purpose, listening construct, task characteristics, and validity evidence. It was found that most of the tests were developed for proficiency and placement purposes in academic contexts, with few of the tests serving for making workplace decisions. Also, global, local, and inferential skills constitute the construct in most listening comprehension tests. A practical approach for justifying the uses of these tests for different stakeholders is discussed. This review is a valuable resource for educators, administrators, test developers, and researchers looking for a comprehensive analysis of existing English tests that assess listening comprehension in second or foreign language learners.

INTRODUCTION

In 2015, more than 3,000 colleges and universities in the United States accepted international students (USA College and University Search, 2015). More than 819,000 international undergraduate and graduate students entered U.S. universities from 2013 to 2014, the highest number ever recorded up to 2014 (Haynie, 2014). To determine the preparedness of international students for universities, prospective international students are often subjected to tests to determine their proficiency in English. Various English language tests have been developed and are used by intensive English programs (IEPs) as well as mainstream academic programs for achievement, placement, and proficiency purposes.

A large number of test reviews on these tests provide useful information to stakeholders (e.g., Alderson, 2009; Fox & Fairbairn, 2011; Kerstjens & Nery, 2000; Macqueen & Harding, 2009; Wang, Choi, Schmidgall, & Bachman, 2012). These test reviews, however, are reviews of individual tests and not an extensive collection of reviews. Stoynoff and Chapelle's (2005) book *ESOL Tests and Testing* filled this gap and reviewed 20 commercially produced English as a second language (L2) tests available at that time. However, in this book, each test was reviewed individually by different authors following the same text organization, namely test

Correspondence concerning this article should be addressed to Tingting Kang, Northern Arizona University E-mail: tk277@nau.edu.

purpose, test methods, and test use justification. Since many useful pieces of information were embedded in the texts, readers may find it hard to quickly compare one feature (e.g., scoring in test methods) across 20 tests. Therefore, a table with a clear classification of each category would facilitate the usefulness of a test review.

Ten years later, most of the tests in Stoynoff and Chapelle's (2005) book have been updated, and various new tests have emerged. In this case, there is a need to review the most recent English as a second language (ESL) and English as a Foreign Language (EFL) academic tests.

Additionally, the test reviews in Stoynoff and Chapelle's (2005) book did not target any specific language skills. Stakeholders who are only interested in assessing learners' listening skills have to extract the information related to listening by themselves after reading all of test reviews. This review, on the other hand, provides a comprehensive summary of each test focused exclusively on the listening construct. This review is particularly useful as listening is an important skill that appears at "the early stages of second language learning" (Nation & Newton, 2009, p. 37), enables learners to receive language input, and facilitates the emergence of other language skills (Vandergrift & Goh, 2012).

This test review was designed to systematically analyze tests used to assess the proficiency of second and foreign language learners in English based on three main aspects: purposes and constructs, characteristics of tasks, and validity evidence. It was important to obtain the psychometric information about the tests and the validity evidence of the tests, providing options of available assessment tools to aid researchers, educators, and administrators in the process of selecting the most suitable test for their purpose. Also, researchers, educators, administrators, and students might find this review useful to be familiar with and informed about existing instruments that measure listening comprehension in a second and foreign language, and how reliable these tests are as part of validity evidence.

The research questions (RQs) guiding this test review were as follows.

RQ 1. What is the purpose and construct of the listening comprehension of the reviewed listening tests?
RQ 2. What are the characteristics of tasks in the reviewed listening tests?
RQ 3. What is the validity evidence of the reviewed listening tests?

METHOD

This section describes the steps that were taken to conduct the review of the listening tests' sections: the identification of the tests, the exclusion criteria, and the analyses.

In an attempt to identify the relevant tests, the 20 tests in *ESOL Tests and Testing* (Stoynoff & Chapelle, 2005) were analyzed first. Seven tests were excluded owing to one or more of the following reasons: a) does not include listening section, b) is no longer available, c) is not for adult learners, and d) does not target academic English. Two online resources, *Mental Measurements Yearbooks* (http://buros.org/mental-measurements-yearbook) and *Google* (https://www.google.com/), were searched, using combinations of keywords (i.e., listening, English, ESL, EFL, test). In addition, several tests on testing company websites (i.e., *Cambridge English Language Assessment, CaMLA*, and *Educational Testing Service*) were analyzed. Seven new tests were added, and a total of 20 tests were reviewed. The following

20 tests met the established criteria of including a listening comprehension section, being currently available, being designed for second or foreign language adult learners of English, and targeting academic language:

ACT ESL Placement Test,
Business Language Testing Service (BULATS),
Cambridge English: Advanced (CAE),
Canadian Academic English Language Assessment (CAEL),
Cambridge Michigan Examination for the Certificate of Proficiency in English (ECPE),
CaMLA English Placement Test—Forms D, E, F (CaMLA EPT),
Certificate of Proficiency in English (CPE),
Comprehensive Adult Student Assessment System (CASAS),
Comprehensive English Language Test (CELT),
Computerized Adaptive Placement Exams (CAPE, or WebCAPE),
EF Standard English Test (EFSET),
First Certificate in English (FCE),
International English Language Testing System (IELTS),
Michigan English Language Assessment Battery (MELAB),
Michigan Test of English Language Proficiency Series (MTELPs),
Pearson Test of English Academic (PTEA),
Test of Adult Basic Education Complete Language Assessment System–English (TABE CLAS-E),
Test of English as a Foreign Language-Internet Based Test (TOEFL iBT),
Test of English for International Communication (TOEIC), and
Woodcock Language Proficiency Battery-Revised (WLPB-R).

After the search for the tests, relevant information on the 20 tests was coded to make a comprehensive analysis of the general information, purposes and listening construct definitions, characteristics of tasks, and validity evidence (see Appendix). Relevant resources of research, articles, sample tests, websites, and other documents that supported the analysis are also listed in the Appendix. If contrasting information was observed, priority has been given to the information from the most updated official websites. The Appendix is a tool that provides specific information about the tests; it can be used as a resource or reference to learn more about each test. Table 1 presents key elements extracted from the Appendix that can help stakeholders have a quick glance over the 20 tests.

RESULTS

This section provides results to address each research question raised in the present study.

Purpose and Construct

RQ1. What is the purpose and construct of the listening comprehension of the reviewed listening tests?

This review focused on the tests of academic English language listening ability of non-native speakers of English in both ESL and EFL contexts. Since most of the tests were academically

TABLE 1
A Quick Glance over the 20 Tests

	Purpose[a]								
Name of the test	*1*	*2*	*3*	*Delivery Mode*[b]	*Testing Approach*[c]	*Time*[d]	*Objectivity*	*Reliability*	*Validity*
1. ACT ESL Placement		✓		CA	C	20	✓	✓	✓
2. BULATS			✓	P & CA	C	50	✓	✓	✓
3. CAE			✓	P & C	C & I	40	✓	✓	✓
4. CAEL			✓	P	C & I	20	✓	✓	✓
5. ECPE			✓	P	C	40	✓	✓	✓
6. CaMLA EPT		✓		P & C	C	60	✓	✓	✓
7. CPE			✓	P & C	C & I	40	✓	✓	✓
8. CASAS	✓			P & C	C & D	50	✓	-	✓
9. CELT		✓		P	C & D	40	✓	✓	✓
10.CAPE		✓	✓	CA	C	25	✓	✓	✓
11. EFSET			✓	CA	C	25-60	✓	-	✓
12. FCE			✓	P & C	C & I	40	✓	✓	✓
13. IELTS			✓	P & C	C & I	30	✓	✓	✓
14. MELAB			✓	P	C	35	✓	✓	✓
15. MTELPs Series	✓			P & C	C	50	✓	-	✓
16. PTEA			✓	C	C & I	45-57	✓	✓	✓
17. TABE CLAS-E			✓	P	C & D	20	✓	-	✓
18. TOEFL iBT			✓	C	C	60-90	✓	✓	✓
19. TOEIC			✓	P & C	C & D	45	✓	✓	✓
20. WLPB-R			✓	C	C, D, & I	-	✓	✓	✓

Note. [a]1: Achievement, 2: Placement, 3: Proficiency; [b]C: computer-based, P: paper-based, & CA: computer-adaptive; [c]C: Communicative testing, D: Discrete-point approach, I: Integrative testing; [d] in minutes

oriented or had the potential to be used in academic settings, their purposes substantially converged. Eighteen of the 20 tests (all except for BULATS and TOEIC) have been primarily designed to assess test-takers' ability to understand academic and general English and their readiness to undertake academic ESL or mainstream studies in English. In other words, the reviewed tests can be used for placement, proficiency, and admission decisions for ESL/EFL programs, undergraduate programs, or higher educational institutions. Even though some of the tests are not primarily intended for these purposes, instead focusing on using English in real-life work situations or business environments, the reviewers considered the designers' claims regarding the possibility of the tests' use for these contexts (n = 4; i.e., BULATS, CAPE, CPE, TOEIC).

The similarity in purpose and intended uses of the tests has made this review particularly valuable in terms of describing the construct of assessment. To some extent, all of the tests share the same overarching construct, which can be termed as academic listening comprehension. However, each test brings its uniqueness with respect to what components have been included in the definition of this construct.

In the literature, various definitions of listening have been offered that differ significantly in scope and formulation. Table 2 summarizes the definitions of listening proposed by different researchers.

TABLE 2
Definitions of Listening Construct

Authors	*Definitions of Listening Construct*
Brown (1949, p. 140)	Get the lecture details, follow a sequence of details in the form of formal directions, keep a series of details in mind until questioned, get the central ideas, draw inferences, distinguish relevant from irrelevant materials, use contextual clues to word meanings, and use transitional elements.
Buck (2001, p. 2)	The process of how nonlinguistic knowledge is applied to the incoming sound: top-down and bottom-up.
Dozer (1997, p. 2)	The process of determining a reason for listening, taking the raw speech and deposits an image of it in short-term memory, organizing the information, predicting information, recalling background information, assigning a meaning to the message, checking that the message has been understood, determining the information to be held in long-term memory, and deleting the original form of the message in short-term memory.
Lund (1990, p. 109)	The process of six listening functions: identification, orientation, main idea comprehension, detail comprehension, full comprehension, and replication.
Lundsteen (1971, p. 9)	The process by which spoken language is converted to meaning in the mind.
Nichols (1947, pp. 83–84)	The attachment of meaning to aural symbols.
Rubin (1995, p. 7)	An active process in which listeners select and interpret information which comes from auditory and visual cues in order to define what is going on and what the speakers are trying to express.
Weir (1993, pp. 51–58)	Direct meaning comprehension; inferred meaning comprehension; contributory meaning comprehension; note taking.
Wolvin (2009, pp. 1–3)	A sequence of behaviors that are generally accepted to characterize the decoding process: receiving, attending, perceiving, interpreting, and responding.
Wolvin and Coakley (1993, pp. 15–22)	Distinguish the auditory and/or visual stimuli; an understanding of the message; provide the speaker the opportunity to talk through a problem; evaluate what is communicated.

None of the reviewed tests explicitly claimed to have followed any specific definition of the listening construct. However, it is evident that all tests are based on some fundamental understanding of listening, which seems to correspond to the definition by Lundsteen (1971) or Nichols (1947) in Table 2. In contrast to having this common concept of listening, the tests differ in the inclusion of particular segments of the listening construct. In this respect, different tests can be related to certain definitions more distinctly, which is discussed below.

Researchers mostly agree on the hierarchical model of listening comprehension, which includes lower- and higher-level skills (Buck, 1991). This taxonomy has been utilized in many tests of L2 listening proficiency and led to the employment of comprehension questions that are believed to assess certain listening subskills, such as understanding main ideas, identifying specific details, and making inferences. This taxonomy is in line with some of the definitions of listening comprehension (i.e., Brown, 1949; Lund, 1990; Weir, 1993). The majority of the tests in this review rely on subskill-oriented items as measures of the bigger construct of academic listening comprehension. Comparatively few tests (n = 4; i.e., CAEL, CaMLA EPT, MTELPs, and CAPE) delimited the construct to no more than the three most widely used listening subskills—understanding main ideas, identifying specific details, and making inferences. In contrast, the majority of the reviewed tests expanded their subconstructs to measure many more subskills, which can be subdivided into the following two categories. One category deals with the understanding of the topic, purpose, organization, and function of the discourse (n = 8; i.e., ACT, CAE, CASAS, CPE, FCE, IELTS, PTEA, and TOEFL iBT). The other category includes the ability to interpret pragmatic aspects of listening comprehension. Thus, in one way or another, ECPE, IELTS, MELAB, PTEA, TABE CLAS-E, and TOEFL iBT (n = 6) measure test-takers' ability to discern speaker's tone, feelings, attitudes, opinions, or intentions. These categories are rarely explicitly included in the existing definitions of listening comprehension. Rather, they are buried under general concepts such as "defined … what the speakers are trying to express" (Rubin, 1995, p. 7) or "assigning a meaning to the message" (Dozer, 1997, p. 2). The tendency to include more specifically defined subskills in the tests may show the development of the listening comprehension construct that now tends to include the ability to comprehend both explicitly and implicitly stated information by employing knowledge from different linguistic subfields, including pragmatics.

More recent tests displayed distinctive features with regard to Buck's (2001) definition of the listening construct. Some tests concentrated on measuring abilities associated with bottom-up listening processes (Buck, 2001; Flowerdew & Miller, 2005), such as comprehending lexical and grammatical meanings (e.g., CASAS, CELT, MELAB), as well as phonological skills (e.g., CASAS). Others included global abilities to synthesize information from different parts of a listening passage, as well as to integrate information from different sources (e.g., ECPE, MELAB, PTEA), which are seldom mentioned as a separate skill to assess in a listening comprehension test.

Based on Buck's (2001) general principle "to include anything that is dependent on linguistic knowledge" (p. 113), the more linguistic aspects of the academic listening comprehension construct are accounted for, the better the construct is represented in an assessment instrument. Building from this, it may be argued that, to fully represent the construct, tests should incorporate numerous subskills, such as those that have been discussed above, as well as innovatively attempt to include other skills that may be related to the overall listening ability.

Some tests (e.g., MELAB) seem to follow the principle to include as many relevant listening subskills as possible. MELAB presents these skills in a clear taxonomy of the listening subconstructs, namely global, local, and inferential skills (see Appendix). Global skills include identifying the main idea, identifying speaker's purpose, and synthesizing ideas from different parts of the stimulus. Local skills consist of identifying supporting detail, understanding vocabulary, synthesizing details, and recognizing restatement. Finally, inferential skills are based on the abilities to understand rhetorical function, make an inference, and make pragmatic implications. This structure allows for meaningful inclusion of explicit and implicit, bottom-up and top-down, as well as informational and inferential approaches to defining and operationalizing a listening comprehension construct.

All of the reviewed tests treated listening comprehension as an exclusively auditory act. No attempt was made to include processing visual information into the construct of listening comprehension. Even though some of the reviewed tests (e.g., ACT, TOEFL) support the incorporation of visuals in the form of pictures, maps, or diagrams, these stimuli are mainly used as pre-planned visual accessories for delivering academic lectures with the purpose of introducing to the scene, activating prior knowledge, or focusing test-takers' attention (Buck, 2001) and have little to do with the ability to interpret unprompted nonverbal signals.

Characteristics of Listening Tasks

RQ2. What are the characteristics of tasks in the reviewed listening tests?

After reviewing the purpose and the listening construct of each test, this RQ focuses on the characteristics of tasks and the listening input. In determining which aspects of task characteristics to include in this article, Dunkel, Henning, and Chaudron's (1993) model of L2 listening comprehension assessment and Buck's (2001) guidelines in designing task characteristics were considered. These aspects include the delivery mode, time of the test, number of times test takers could hear the input, task structure, English varieties, and scoring method. The purpose of this section is to provide a general overview and the current trends associated with each of these characteristics. Detailed information of each test is presented in the third column in the Appendix.

The delivery mode can be classified into paper-based and computer-based. With the advancement of technology, there has been an increasing number of computer-based English language assessments (n = 16; i.e., ACT ESL Placement Test, BULATS, CAE, CaMLA EPT, CPE, CASAS, CAPE, EFSET, FCE, IELTS, MELAB, MTELPs, PTEA, TOEFL iBT, TOEIC, and WLPB-R) because of its main advantage in terms of test taking and scoring. Another noticeable trend is the use of a computer-adaptive approach, or a "value-added approach to computerization of the paper-and-pencil format" (Jamieson, 2005, p. 230), which provides shorter timed tests and adjusts the level of difficulty of test items to suit test-takers' levels of proficiency. The computer-adaptive listening tests include ACT ESL Placement test, BULATS, CAPE, and EFSET (n = 4). However, the paper-and-pencil counterpart still coexists. Some publishers offer both modes of delivery as alternative options, such as in CAE, CaMLA, and IELTS. Others provide only paper-based listening tests, CAEL, ECPE, CELT, and MELAB (n = 4).

In general, the time of taking the tests ranges from 20 to 90 minutes. The tests that are time-efficient and can be completed in 20 minutes have different modes of delivery. For computer-adaptive tests, test takers of ACT ESL Placement Tests, CAPE, and EFSET can complete the tests within 20 or 25 minutes. WLPB-R allows test takers to finish in 20 minutes although the test can take up to 60 minutes. The paper-based tests including CAEL and TABE CLAS-E also require approximately 20 minutes of test-taking time. The test that requires the longest amount of time to finish is TOEFL iBT (60–90 minutes).

Thirteen tests allow test takers to hear the listening input only once. In BULATS, CAE, CPE, and FCE, the input can be heard twice. Interestingly, CASAS varies the number of hearings according to test-takers' levels of proficiency. The tendency of retaining only one hearing suggests that most listening tests are in favor of measuring automatic processing of test takers. As Buck (2001) described, automatic language processing is important because in real-world situations we hear the input only once and fast. If testtakers fail to understand the message when the input is played only once, the ability to inference will help bridge the gaps. Therefore, this listening ability is what test developers want to measure.

In assessing listening, three main approaches have been used to operationalize the listening construct, namely, discrete-point approach, integrative testing, and communicative testing (Buck, 2001). The most common tasks for testing L2 listening in the discrete-point tradition are "phonemic discrimination tasks, paraphrase recognition, and response evaluation" (Buck, 2001, p. 63). Dictation, translation, and cloze tests are examples of integrative testing. The communicative approach highlights the use of the language for its communicative functions. Although a test format of communicative approach follows the discrete-point approach by using a multiple-choice item, its primary goal is for overall comprehension and communicative purposes. The academic listening construct has been operationalized differently by the test publishers; however, all 20 tests tend to have a primary focus on communicative purpose (see Table 1). All 20 tests emphasize how language is used in the target domain (i.e., academic and workplace settings). Although listening input is made from a semiscripted text where there are hesitations and fillers, it is designed to reflect many aspects of the real world. Some tests including CAE, CAEL, CPE, FCE, IELTS, PTEA, and WLPB-R (n = 7) incorporate integrative testing in which other skills in processing language have been integrated in the tests. In these tests, the test developers either use cloze or gap-filling summaries to assess L2 test takers. Five tests (i.e., CASAS, CELT, TABE CLAS-E, TOEIC, and WLPB-R) seem to follow the traditional discrete-point approach in which the units of language are measured. For example, on CASAS test takers have to be able to discern different sounds through a minimal pair task, and test takers on TOEIC are asked to listen to a question and then evaluate and choose the best response to that question.

In terms of task structures, the general trend of these tests shows that the listening input on the tests varies in length ranging from short conversations or themed monologues to longer discourse, such as an academic lecture. In computer-adaptive tests, such as ACT ESL Placement test and BULATS, the length of the listening input is adjusted according to test-takers' performance. One noteworthy aspect is that, among the 20 tests, CAEL is the only test that is fully integrated and topic-based. In other words, all the tasks across reading, listening, and writing sections of the test share the same theme. In addition, most listening tests attempt to replicate the real world language listening tasks in various settings at home, school, or workplace. Furthermore, it can be seen that in some tests (e.g.,

TOEFL and IELTS) the listening input is delivered along with a pictorial or graphic display. These visuals provide the test takers with contextual cues as well as increase the authenticity of the test.

Another aspect of the task structure is the response type. Most of the tests require test takers to select the correct answer from either three- or four-option multiple choices, which is practical in terms of ease for scoring. In some tests, such as the TOEFL iBT, test takers have to choose two correct answers in order to receive full credit for the item. Other response types that are included in the tests consist of multiple matching, short answers, filling in the blanks, completing tables or flow charts, note-taking, and arranging event or steps.

As there have been concerns about whether the listening input should reflect a real target language use domain by incorporating a range of accent varieties in listening tests, test developers have considered expanding the scope of accents beyond native varieties of English (Major, Fitzmaurice, Bunta, & Balasubramanian, 2002, 2005). However, Major et al.,, 2005) found that ethnic and international dialects might pose more difficulties to L2 test takers than regional dialects. In this realm, the accents of speakers in the reviewed listening tests represent a variety of speakers from English speaking countries although most are still restricted to United Kingdom, North America, and Australia as reported by the test publishers (BULATS, CPE, FCE, IELTS, PTEA, TOEFL, and TOEIC). Only IELTS and TOEFL include a New Zealand accent and TOEIC reports the inclusion of a Canadian accent. Still, considering the increasing number of international professors, teaching assistants, and students in academic settings, test publishers should acknowledge this reality. Having a non-native variety in listening tests may be a promising start; however, future research studies are needed in order to determine the extent to which a non-native accent would not favor L2 test takers of that particular accent (Harding, 2012).

The scoring methods employed on the tests can be categorized into two broad types: dichotomous and partial scoring. Most of the test tasks are scored dichotomously, and then raw scores are converted to scale scores as seen in FCE, CASAS, MELAB, and TOEFL iBT. There are only a few listening tests that permit partial scoring, such as CAEL and PTEA. For the tests that use computer-adaptive approach, as the ACT ESL Placement Test does, the adaptive scoring method, which takes into account such aspects as difficulty of items and their discriminating power, is implemented accordingly.

With regards to the score report, most of the listening tests (n = 8; i.e., BULATS, CaMLA EPT, ECPE, EFSET, MELAB, MTELPs, PTEA, and TABE CLAS-E) are aligned with the Common European Framework of Reference for Language (CEFR) which describes language proficiency in terms of six ascending tiers, namely A1 and A2, B1 and B2, and C1 and C2. Despite the strengths of the CEFR, one shortcoming should be addressed. Evidence and guidelines of how to make the alignment of the CEFR and other assessments are still in need (Deville, 2012).

Validity Evidence

RQ3. What is the validity evidence of the reviewed listening tests?

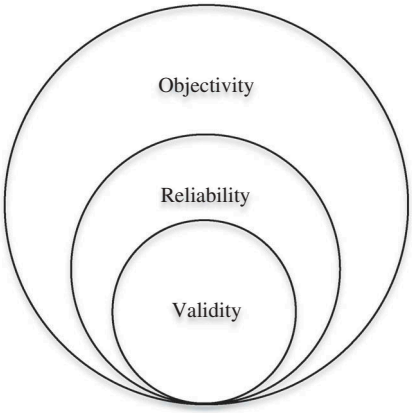

FIGURE 1 Relationship among objectivity, reliability, and validity.

Psychometric standards which need to be met are objectivity, reliability, and validity (Miller, Linn, & Gronlund, 2009). As depicted in Figure 1, objectivity is one of the major factors that influence reliability measures, and reliability is a type of evidence for validity. To reach the ultimate goal of a valid test, the test should be objective and reliable. This section condenses information from the reviewed tests regarding objectivity, reliability, and validity.

The objectivity of a test means "the degree to which equally competent scores obtain the same results" (Miller et al., 2009, p. 126). Two essential ways to ensure test objectivity are standardized instructions and scoring instructions. As shown in the fourth column in the Appendix, all 20 tests have standardized instructions. The listening instructions are scripted and given by trained proctors or the instructions are prerecorded and played automatically. With regards to the scoring instructions, all 20 tests include objective item types, such as multiple choice and matching. These types of items are machine scored. There are also some tests that encompass subjective item types (n = 3; i.e., CAEL, PTEA, TOEFL iBT). With the PTEA test, for example, each item was rated independently by two trained raters to achieve the objectivity and the accuracy of the rater-awarded scores. If there was a disagreement between the two independent raters, a third rating was gathered and the two closest ratings were retained.

According to Miller et al. (2009), "reliability refers to the consistency of measurement, that is, how consistent test scores or other assessments results are from one measurement to the other" (p. 107). In testing, particularly in standardized testing, reliability provides certainty that the results of a test can be trusted, and that it provides the information needed to make decisions such as certification, placement, and achievement.

There are several methods to estimate reliability: test-retest, equivalent-forms, test-retest with equivalent forms, split half, coefficient alpha, and interrater (Miller et al., 2009). These methods measure stability, equivalence, internal consistency, and consistency of ratings. Testing companies report reliability estimates through internal consistency (e.g., Cronbach's alpha, Kuder-Richardson Formula 20 (KR-20), Rasch reliability, and marginal reliability), consistency of ratings (e.g., Spearman rank order correlation), stability (e.g., Pearson

TABLE 3
Reliability Estimates of All of the 20 Tests

Name of the test	*Method*	*Reliability Estimates*
1. ACT ESL Placement	Cronbach's alpha	.85-.89
2. BULATS	Rasch reliability	.92
3. CAE	Cronbach's alpha	.73
4. CAEL	Spearman rank-order correlation coefficient	.97
5. ECPE	KR-20	.83
6. CaMLA EPT	Cronbach's alpha	.98
7. CPE	Cronbach's alpha	.74
8. CASAS	-	-
9. CELT	KR-20	.88-.96 (overall test; did not provide listening section's reliability)
10.CAPE	Test-retest	.76-.86
11. EFSET	-	-
12. FCE	Cronbach's alpha	.81
13. IELTS	Cronbach's alpha	.90
14. MELAB	KR-20	.88
15. MTELPs Series	-	-
16. PTEA	Split-half	.89
17. TABE CLAS-E	-	-
18. TOEFL iBT	Cronbach's alpha	.85
19. TOEIC	Cronbach's alpha	.92-.93
20. WLPB-R	Cluster reliability	.90 s (overall test; did not provide listening section's reliability)

correlation), and equivalency (e.g., equating). Regardless of the method, reliability is always reported on a scale from zero to one. For high–stakes, professionally developed tests, a reliability index is expected to be close to or greater than .90. Table 3 includes the reliability estimates observed from the 20 tests.

Generally speaking, reliability coefficients seem to have reached the necessary coefficients for high-stakes tests. The highest coefficient is .98, observed from CaMLA EPT. It is noteworthy that the lowest coefficients of the listening sections of these standardized tests are in the .70 s for CAE and CPE. Reliability coefficients are not available for CASAS, MTELPs, TABE CLAS-E, and EFSET. The first three tests (i.e., CASAS, MTELPs, and TABE CLAS-E) are purchased by institutions in the forms of books or packages. The data that result from test administrations might not be available for testing companies to conduct reliability analyses. The last test, EFSET, is a free online adaptive test that, according to the technical report, conducts reliability analyses using Item Response Theory; however, the actual coefficients were not available.

Nowadays, reporting reliability seems to be a box to check when presenting psychometric information on a test but the information from the coefficients may be left unexplained. The quality of reliability in tests can be considered a part of the validity evidence of a test. Chapelle (1999) defined validity as an argument "concerning test interpretation and use: the extent to which test interpretations and uses can be justified" and based on "the basis of a number of types

of rationales and evidence, including the consequences of testing" (p. 258). Specifically, construct validation is widely accepted as central to validity since the 1980s (e.g., Chapelle, 1999; Kane, 2012; Messick, 1980). More recently, Kane (2012) developed the argument-based approach to validation, which involves two steps: "First, specify the proposed interpretations and uses of the scores in some detail; second, evaluate the overall plausibility of the proposed interpretations and uses" (p. 4). The interpretations and uses of the scores have been discussed in the previous *Purpose and Construct* section, where test purposes and listening constructs have been clearly identified by different test developers. In terms of the second step in Kane's (2012) argument-based approach, various construct validity evidence has been proposed, such as content analysis, empirical item or task analysis, dimensionality analysis, investigation of relationships of test scores with other tests and behaviors, analysis of different test performance, and testing consequence analysis (Chapelle, 1999; Messick, 1989). Owing to the broadness and abstractness of validation, the validity evidence presented by the testing companies varied. However, to obtain strong conclusions, the more evidence test developers provide, the more confidence will be achieved in test validation. Following are some trends that have been observed in the validity evidence section.

First, many of the reviewed tests (n = 9; i.e., BULATS, CaMLA EPT, ECPE, EFSET, MELAB, MTELPs, PTEA, TABE CLAS-E, and TOEIC) offer a range of results that have been mapped to the CEFR. Moreover, ECPE has been designed to target specific levels of the CEFR by designing items with the subskills (e.g., listening for a gist and interpreting meaning) using the descriptors of ability (e.g., B1 and C1), language functions (e.g., talking about familiar topics), and contexts of language use (e.g., academic and social) as the source for test development.

Second, some of the test validations focus on the content analysis (n = 3; i.e., ACT ESL Placement Test, PTEA, and TOEFL iBT) and empirical item analysis (n = 4; i.e., ACT ESL Placement Test, MELAB, PTEA, and TOEIC). For example, in TOEFL iBT, the content of the test is relevant to and representative of the kinds of tasks and written and oral input that students encounter in college and university settings, and all items in ETEA undergo a series of review process.

Another source of validity is criterion validity through correlating the test scores to other test versions or well established tests (n = 3; i.e., CaMLA EPT, CELT, and MELAB). For instance, CELT reported strong correlations with TOEFL (r = .79) and MTELPs (r = .81).

Generally, testing companies such as *Educational Testing Service* and *Cambridge English Language Assessment* devote part of their efforts to promote research. In fact, researchers conduct studies that contribute to the validity evidence of their tests using data that has been collected during examinations. Previous to launching in 2013, the CaMLA EPT-Forms D, E, F (paper-pencil and computer based) were piloted with 480 test takers with different first language backgrounds (Walter & Hentschel, 2013). As part of the test development process, four forms were piloted with 91 new items and 15 items from Form A to conduct equating of the old and new forms. In the end, all the forms of the CaMLA EPT test are comparable.

Additionally, cut-off score analysis, fairness check, and consequence analysis are only mentioned by a few tests (n = 4; i.e., ACT ESL Placement, CAEL, CaMLA EPT, and MTELPs Series). The lack of information related to these topics does not allow us to fully capture the validity of the tests.

To summarize, various test validation evidence have been provided by different test developers. However, some of them are incomplete and lack an explanation. Following a comprehensive validity evidence framework (e.g., Chapelle, 1999; Messick, 1989) could establish a more systematic analysis in test validation.

DISCUSSION

This study critically reviewed 20 academic ESL/EFL listening tests in respect to their purpose, operational definitions of the listening construct, task characteristics, and overall validity evidence. Substantial difference among tests were found in each area, which may lead to different degrees of usefulness and appropriateness of the tests.

To justify the uses of these tests for different stakeholders, in the following section a practical approach in evaluating the test usefulness, proposed by Bachman and Palmer (1996), is discussed. In determining whether a particular test is appropriate to your specific context, six qualities of test usefulness should be taken into account: reliability, construct validity, authenticity, interactiveness, impact, and practicality (Bachman & Palmer, 1996). After identifying the main purpose of why a listening test is needed (i.e., for achievement, placement, or proficiency), the first and foremost step in evaluating the test is the *reliability*, or to what extent the test gives consistent scores over time. Most of the listening tests reviewed here report their reliability estimates; however, the reliability of four tests including CASAS, EFSET, MTELPs Series, and TABE CLAS-E could not be found.

Next, the *construct validity*, or whether the test scores can be used to make inferences about the construct. As discussed in the previous section, the listening tests share similar underlying construct: academic listening comprehension. However, the constituents of the construct vary across the tests. Therefore, a program evaluation is recommended so that the chosen test is parallel to the overall content of the courses. In addition, it is advised that test stakeholders should consider how fully the construct of a particular instrument is represented in terms of different aspects of assessing listening ability. Choosing a test that will assess a reasonably large number of listening subskills is the next step. Also, it seems worth looking for a test that would incorporate visual cues along with auditory input, which will increase the authenticity of a listening assessment instrument (Buck, 2001) as well as, possibly, the construct validity of the instrument.

Most test tasks are designed to capture language use in real world contexts; thus, the *authenticity* of the test also needs to be considered. Apart from representing the real academic contexts, several tests also correspond with the business and professional settings, including BULATS, CASAS, CAPE, FCE, IELTS, MELAB, TOEFL iBT, and TOEIC Listening & Reading Test. Still, the authenticity of the reviewed tests is largely suffering by not assessing the ability to interpret visual cues as part of a listening process.

In addition, when taking the test, test takers engage in using "language knowledge, metacognitive strategies, topical knowledge, and affective schemata" (Bachman & Palmer, 1996, p. 25); in other words, the test has *interactiveness*. As we are limited to full access of the real tests, it is advisable to examine the language input and determine the extent of the involvement of the task takers' individual characteristics. Such evaluation questions could be: "To what extent does the task presuppose the appropriate area or level of topical

knowledge? Does the processing required in the test task involve a very narrow range or a wide range of areas of language knowledge? How much opportunity for strategy involvement in provided? Is this test task likely to evoke an affective response that would make it relatively easy or difficult for the test takers to perform at their best" (Bachman & Palmer, pp. 142–145).

Furthermore, the test should have positive *impact* or consequences, to all stakeholders; for instance, the test takers should be able to use the scores as a valid indicator of what they know, and at the same time the teachers can use the overall scores to evaluate their teaching.

The last factor to consider is *practicality*, which concerns logistic issues within institutions. The cost of implementing the test may be one of the major concerns because most listening tests cannot be separately taken out from other language skills: speaking, reading, and writing, except for CASAS. The delivery mode (e.g., paper- or computer-based and test-taking time) should also be considered in choosing which test would best suit a particular setting. For example, implementing the computer-adaptive CAPE in institutions where the facilities are limited may not be practical.

LIMITATIONS AND FUTURE DIRECTIONS

This test review has some limitations. First, due to the accessibility of the tests, some of the test information is not available for further review (e.g., actual test, reliability report, detailed validity evidence). Therefore, all the obtained results are limited to online public sources. Second, the reviewed test-takers' population is restricted to adult English language learners. In the future, both age group and target language could be expanded following the same test review format. For instance, listening tests for younger learners and other languages, such as Arabic, Chinese, and Spanish, could all be added to this test list.

From the present review, one noticeable gap in operationalizing the academic listening comprehension construct emerges. As some researchers have argued (Ockey, 2007; Rubin, 1995; Sueyoshi & Hardison, 2005; Wagner, 2008; Wolvin & Coakley, 1993), the interpretation of nonverbal information such as gestures, facial expression, and body movement and posture, which are found in the absolute majority of authentic listening interactions, can be considered a part of the construct definition of the L2 ability. This notion deals with the inclusion of visual stimuli into the listening tests. Some of the reviewed tests incorporated pictorial stimuli (i.e., ACT, TOEFL) along with the listening input. However, these tests did not address the potential of using visual stimuli for constructing the meaning of a listening message. Visuals used in the tests were not to explain or illustrate the answers to listening comprehension questions, which is the opposite to what often happens in authentic academic listening situations. It seems that none of the reviewed tests have attempted to assess test-takers' ability to interpret visual cues as such.

Future research could attempt to investigate if the inclusion of visual information into medium- and high-stakes tests would enhance the usefulness of the tests according to Bachman and Palmer's (1996) model. Conducting quantitative validation studies as well as gathering qualitative information (e.g., opinions and attitudes) from test takers, test

developers, and other people in the field of second or foreign language teaching would be one way to do this.

Another suggestion is to review other second or foreign language high-stakes listening assessments that include visual information in some form or claim to test the ability to interpret it. Such a review would provide additional insights into how well the "visuals-inclusive" approach lends itself to real-life testing situations and identify the effects for test validity.

PRACTICAL IMPLICATIONS

The academic listening tests included in this study were reviewed based on the test purpose and construct, the characteristics of the tasks, and the validity evidence. Reviewing these 20 tests provides the most updated available listening tests in the field of English language assessment. Among the most relevant results, proficiency and placement test by far constitute the majority of the available tests. Moreover, global, local, and inferential skills make up the construct of listening comprehension in most of the tests. Regarding the characteristics of the listening tasks, an emergent delivery mode using computers can be observed. Finally, different types of reliability indices are provided by the testing companies; however, the validity argument of the tests would be strongly supported by having more extensive evidence.

This test review includes detailed and organized information that will be useful to second language teachers, program administrators, and test developers. Depending on the purpose, listening subconstructs measured, time allowed, type of the questions, characteristics of the input message, and the validity of the tests, L2 teachers and program administrators can make an informed decision in choosing the listening test that is best suited to their institutions. Test developers can also compare and contrast these tests to refine their own tests. The Appendix is a good reference tool for researchers because the table condenses information about each test. Additionally, references provided in end could be used for further research.

ACKNOWLEDGMENTS

The authors thank Dr. Joan Jamieson for her valuable comments and encouragement for this review. Our sincere thanks to anonymous reviewers and the editors for their constructive comments and suggestions. We are also thankful to Jacqueline Church for proofreading the manuscript.

REFERENCES

Alderson, J. C. (2009). Test review: Test of English as a foreign language™: Internet-based test (TOEFL iBT®). *Language Testing*, *26*, 621–631. doi:10.1177/0265532209346371

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. New York, NY: Oxford University Press.

Brown, J. I. (1949). The construction of a diagnostic test of listening comprehension. *The Journal of Experimental Education*, *18*, 139–146. doi:10.1080/00220973.1949.11010402

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, *8*, 67–91. doi:10.1177/026553229100800105

Buck, G. (2001). *Assessing listening*. New York, NY: Cambridge University Press.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272. doi:10.1017/S0267190599190135

Deville, C. (2012). Book review: Aligning tests with the CEFR: Reflections on using the council of Europe's draft manual. *Language Testing*, *29*, 312–314. doi:10.1177/0265532211434013

Dozer, C. V. (1997). *Improving ESL learners' listening skills: At the workplace and beyond* [PDF]. Retrieved from http://www.marshalladulteducation.org/pdf/briefs/Improving%20ESL%20Lrns%20List.vanDuzerdoc.pdf

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *The Modern Language Journal*, *77*, 180–191. doi:10.1111/modl.1993.77.issue-2

Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice*. New York, NY: Cambridge University Press.

Fox, J., & Fairbairn, S. (2011). Test review: ACCESS for ELLs®. *Language Testing*, *28*, 425–431. doi:10.1177/0265532211404195

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, *29*, 163–180. doi:10.1177/0265532211421161

Haynie, D. (2014, July 29). International students flock to these universities. *U.S. News*. Retrieved from http://www.usnews.com/education/best-colleges/slideshows/us-universities-with-the-most-international-students

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, *25*, 228–242. doi:10.1017/S0267190505000127

Kane, M. (2012). Validating score interpretations and uses: Messick lecture, language testing research colloquium, Cambridge, April 2010. *Language Testing*, *29*, 3–17. doi:10.1177/0265532211417210

Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. *IELTS Research Reports*, *3*, 85–108.

Lund, R. J. (1990). A taxonomy for teaching second language listening. *Foreign Language Annals*, *23*, 105–115. doi:10.1111/flan.1990.23.issue-2

Lundsteen, S. W. (1971). *Listening: Its impact on reading and the other language arts*. Urbana, IL: National Council of Teachers of English.

Macqueen, S., & Harding, L. (2009). Review of the Certificate of Proficiency in English (CPE) speaking test [Review of the test The Certificate of Proficiency in English (CPE) speaking test]. *Language Testing*, *26*, 467–475. doi:10.1177/0265532209104671

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, *36*, 173–190. doi:10.2307/3588329

Major, R. C., Fitzmaurice, S. M., Bunta, F., & Balasubramanian, C. (2005). Testing the effects of regional, ethnic, and international dialects of English on listening comprehension. *Language Learning*, *55*, 37–69. doi:10.1111/j.0023-8333.2005.00289.x

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027. doi:10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Miller, M. D., Linn, R., & Gronlund, N. (2009). *Measurement and evaluation in teaching* (10th ed.). Upper Saddle River, NJ: Merrill, Prentice Hall.

Nation, I. S. P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking*. New York, NY: Routledge.

Nichols, R. G. (1947). Listening: Questions and problems. *Quarterly Journal of Speech*, *33*, 83–86. doi:10.1080/00335634709381268

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*, 517–537. doi:10.1177/0265532207080771

Rubin, J. (1995). An overview to 'A guide for the teaching of second language listening'. In D. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 7–11). San Diego, CA: Dominie Press.

Stoynoff, S., & Chapelle, C. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*, 661–699. doi:10.1111/j.0023-8333.2005.00320.x

USA college and university search: International student [Website]. (2015). Retrieved from http://www.internationalstudent.com/school-search/usa/

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. New York, NY: Routledge.

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, *5*, 218–243. doi:10.1080/15434300802213015

Walter, D., & Hentschel, J. (2013). *CaMLA English Placement Test (EPT) forms D-F: Development report* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/EPT-Development-Report-20131.pdf

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*, 603–619. doi:10.1177/0265532212448619

Weir, C. (1993). *Understanding and developing language tests*. New York, NY: Prentice Hall.

Wolvin, A. D. (2009). Listening, understanding, and misunderstanding. *21st century communication: A reference handbook*, *1*, 1–15.

Wolvin, A. D., & Coakley, C. G. (1993). A listening taxonomy. In A. D. Wolvin & C. G. Coakley (Eds.), *Perspectives on listening* (pp. 15–22). Norwood, NJ: Ablex.

APPENDIX

Academic Listening Test Reviews

This table condenses information on the Listening section of 20 academic English tests. The main focus of analysis is the Listening section of the tests. References found at the end are keyed by numeric superscripts. The following information is within each column:

TEST NAME: A. publisher, b. year of publication, c. website, d. cost

PURPOSES & LISTENING CONSTRUCTS: A. intended use, b. description of listening constructs

CHARACTERISTICS OF LISTENING TASKS: A. delivery mode, b. time of the listening section, c. number of hearing, d. task structure, e. English accents, f. scoring method

VALIDITY EVIDENCE: A. objectivity, b. reliability, c. validity justification

TEST NAME	*PURPOSES & LISTENING CONSTRUCTS*	*CHARACTERISTICS OF LISTENING TASKS*	*VALIDITY EVIDENCE*	*COMMENTS*
1. ACT ESL Placement Test a. Act, Inc.[1] b. 1999[2] c. http://www.act.org/compass/tests/esl.html d. $1.55 U.S. dollars ($US) per unit (the institution decides on the test package that may require multiple units)[1]	a. Placement test. It is used to help postsecondary institutions quickly and accurately assess incoming ESL students' English language ability levels and place them into appropriate ESL courses.[3] b. The abilities to recognize and manipulate Standard American English in two main categories: listening for explicitly stated information: recognizing main ideas, significant details, and relationships; determining sequence and relationship from discourse markers; recognizing numbers and dates; listening for implicitly stated information: understanding main ideas; making inferences about omitted information; determining vocabulary from context clues; recognizing register.[3]	Computer-adaptive[3] b. Untimed but typically 20 minutes[1] c. Heard once[3] d. Aural stimuli adaptively differing in length and complexity by proficiency level: sentence-length prompts and short conversation exchanges (beginners), longer prompts of primary everyday experiences (intermediate), more complex conversational exchanges and short academic lectures (advanced); Four-option multiple-choice items; pictorial or textual with the facility to listen to recordings of questions[3] e. American accents[1] f. Taking into account not only the number of correct answers, but also the difficulty of items, the typical probability of guessing, and the discriminative power of the items[3]	a. Objectivity: standardized instruction - the instructions are pre-recorded and played automatically. Scoring instructions: machine scoring.[4] b. Reported marginal reliability coefficient was .85 - .89 obtained through simulation studies for the listening section.[3] c. Validity evidence: validation of cut-off scores, interpretation of scores, relevance of test content (selection and training of items writers, items reviews, soundness reviews, sensitivity reviews), relevance of proficiency descriptors, fairness check, differential item functioning analysis[3]	Medium-stakes test[1] Scores ranging from 1 to 99 discriminating test takers among 5 levels of proficiency[3]

2. Business Language Testing Service (BULATS) a. Business Language Testing Service, University of Cambridge Local Examinations Syndicate (UCLES)[1] b. 1997–1998 (paper-based), 2000 (computer-based)[2] c. http://www.bulats.org d. Around $40 Hong Kong dollars[2]	a. Proficiency test. It is used internationally for business and industry recruitment, for identifying and delivering training, for admission to study business-related courses and for assessing the effectiveness of language courses and training.[3] b. The ability to understand the language in a workplace: understanding short conversations; taking down phone messages and notes; listening for gist; identifying topic, context, and function; listening to extended speech for detail and inference.[4]	a. Paper-based or computer-adaptive (online or CD-ROM)[4] b. Approximately 50 minutes[4] c. Heard twice[4] d. Part 1 (short conversations or monologues): 10 discrete three-option text or graphic multiple-choice items; Part 2 (phone messages orders, notes, etc.): 12 form-filling items; Part 3: (short monologues or dialogues), 10 nine-option multiple matching items; Part 4 (extended monologue or dialogue): 18 three-option multiple choice items[4] (Note. For a computer-based test, three-option multiple-choice items only, excluding taking down phone messages) e. A variety of native speaker accents including British, American, and Australian[5] f. Using adaptive testing technique; rapid results and instant test report	a. Objectivity: standardized instruction - the instructions are scripted and given by trained proctors;[7] the instructions are pre-recorded and played automatically.[4] Scoring instructions: machine scoring.[1] b. Reported Rasch reliability indices were .94 for overall computer based-test and .92 for computer-based listening section.[6] c. Validity evidence: Authentic workplace situations to test a candidate's ability to use language in real business situations[3] Scores are aligned with CEFR.[4] All Cambridge systems and processes for designing, developing and delivering exams and assessment services are certified as meeting the internationally recognized standard for quality management.[7]	No requirement of previous business experience and available in four languages – English, French, German, and Spanish[4] A scale of 1-100, no "pass" score; discriminating test takers among 6 levels of proficiency[4]

(*Continued*)

(Continued)

TEST NAME	*PURPOSES & LISTENING CONSTRUCTS*	*CHARACTERISTICS OF LISTENING TASKS*	*VALIDITY EVIDENCE*	*COMMENTS*
3. Cambridge English: Advanced (CAE) a. Cambridge English Language Assessment[1] b. 1991; regular updates; last update in 2015[2] c. http://www.cambridgeenglish.org/exams/advanced/ d. Fees are set by test centers[3], approximately 150 euros.	Proficiency test. Its is used for high achievers who want to: follow an academic course at university level, communicate effectively at managerial and professional level, participate with confidence in workplace meetings or academic tutorials and seminars, carry out complex and challenging research, and stand out and differentiate themselves.[2] The ability to follow monologues (announcements, radio broadcasts, speeches, talks, lectures, anecdotes, etc.) and interacting speakers (radio broadcasts, interviews, discussions, etc.) and understand: the main points of a text, specific information, and stated opinion; gist, detail, function, agreement, and course of action; speaker's purpose, feelings, attitudes and opinions[2]	a. Paper-based or computer-based[2] b. Approximately 40 minutes[2] c. Heard twice[2] d. Part 1 (3 short unrelated dialogues): 2 multiple-choice items on each extract; Part 2 (3-minute monologue): 8 sentence completion items; Part 3 (4-minute conversation between two or more speakers): 6 four-option multiple-choice items; Part 4 (5 short 30-second themed monologues): 2 parallel eight-option multiple-matching tasks with different focus[2] e. A variety of native speaker accents including British, American, and Australian[2] f. Exact match scoring (1 point for each correct answer); the Statement of Results containing scores on Cambridge English Scale (from 80 to 230) for each section as well as overall score[2]	a. Objectivity: standardized instruction - the instructions are scripted and given by trained proctors OR the instructions are pre-recorded and played automatically depending on the delivery mode.[2] Scoring instructions: machine scoring OR scores are given by trained raters depending on the delivery mode.[1] b. Reported reliability coefficients (Cronbach's Alpha) were .93 (internal consistency) with SEM of 2.89 for the total score and 0.73 (internal consistency) with SEM of 2.33 of the listening section.[6] c. Validity evidence: all Cambridge systems and processes for designing, developing and delivering exams and assessment services are certified as meeting the internationally recognized standard for quality management.[5]	Targeted at Level C1 – the second highest level on the CEFR scale[2] Performance ranging between CEFR Levels B2 and C2 also receiving a certificate[2] Results available online for free and verified by institutions instantly[4]

4. Canadian Academic English Language Assessment (CAEL) a. Carleton University, The CAEL Assessment[1] b. 1989[2] c. https://www.cael.ca d. $185 Canadian dollars[3]	a. Proficiency test. It is used to describe the level of English language of test takers planning to study in English-medium colleges and universities.[3] b. The ability to listen to, take notes, and transfer or apply information on a topic introduced or extended by an academic lecture, which includes: comprehending main ideas and comprehending specific details.[3]	a. Paper-based[3] b. Approximately 20 minutes[3] c. Heard once[3] d. One academic lecture themed with the reading section and followed by comprehension questions of different types: short answer, filling in the blanks, completing tables, multiple choice, note taking, and transferring information to a flow-chart[3] e. Not provided f. Conducted by trained raters using detailed marking key; partial credit permitted on all of the questions	a. Objectivity: standardized instruction - the instructions are scripted and given by trained proctors. Scoring instructions: scores are given by trained raters.[3] b. Reported inter-rater reliability in Spearman rank-order correlation coefficient was .95 (n=178) of the whole test and .97 (n=178) of the listening section.[3] c. Validity evidence: construct representation, construct irrelevant variance, criterion-related evidence of validity, and consequences of test use.[3]	Fully integrated and topic-based (same topic for every subtest)[1] The information from the lecture, as well as from the reading sections of the CAEL synthesized by test takers in the essay for the writing section of the test[3] The band scores ranging from the Very Limited Listener (10-20) to the Expert Listener (80-90)[3]
5. Cambridge Michigan Examination for the Certificate of Proficiency in English (ECPE) a. Cambridge Michigan Language Assessments b. 1953[1] c. http://www.cambridgemichigan.org/institutions/products-services/tests/proficiency-certification/ecpe/ d. US$40.00[2]	a. Proficiency test. It is used to certify EFL adult learners. It targets a C2 level of the CEFR.[3] b. The ability to master three types of questions: global, local, and inferential: Global-Understanding main idea, identifying speaker's purpose, synthesizing ideas from different parts of the text; Local-Identifying supporting detail, understanding vocabulary, synthesizing details, recognizing restatement; Inferential-Understanding rhetorical function, making an inference, inferring, supporting detail, understanding pragmatic implications.[3]	a. Paper-based[4] b. Approximately 35-40 minutes[4] c. Heard once[4] d. Part 1 (dialogues); Part 2 (monologues); Part 3 (dialogues) with tasks around public, occupational, and educational topics: 50 multiple-choice items[4] e. American accent[4] f. A scale score ranging from 0–1000[3]	a. Objectivity: Standardized instructions in testing centers. Scoring instructions: machine scoring[3] b. Reported reliability of the listening section was .83 (internal consistency) using Kuder-Richardson Formula 20 (KR-20) with SEM estimate of 48.66.[5] c. Validity evidence: This test targets mastery in level C2 of the CEFR.[3] Speaking in this test was investigated by Lin (2015).[6]	Administered twice a year in CaMLA centers Note-taking allowed during listening Scores reported in terms of proficiency: Honors (840-1000), Pass (750-835), Low Pass (650-745), Borderline Fail (610-645), Fail (0-605)[3]

(*Continued*)

(Continued)

TEST NAME	PURPOSES & LISTENING CONSTRUCTS	CHARACTERISTICS OF LISTENING TASKS	VALIDITY EVIDENCE	COMMENTS
6. CaMLA English Placement Test - Forms D, E, F (CaMLA EPT) a. Cambridge Michigan Language Assessments b. 2013[1] c. http://www.cambridgemichigan.org/institutions/products-services/tests/placement-progress/camla-ept/ d. Paper and Pencil packages (20 booklets of each form D, E, F, scoring stencils, CDs, examiner manual, reproducible answer sheet) are available for institutions at US$845. For 25 CB EPT the cost is US$148.75.[2]	a. Placement test. It used for placement purposes to assesses listening comprehension, grammatical knowledge, vocabulary range, and reading comprehension for placement purposes. This test ranges from beginner to advanced levels. Institutions around the world use it and adapt it to their placement needs. This is a new version of the EPT developed in 1976 (Forms A, B, C).[1] b. The ability to understand main idea, inferencing, details, responding to a question or statement[1,3]	a. Paper-based and Computer based (available in 2015); Forms D, E, and F[4] b. Approximately 60 minutes[3] c. Heard once[1] d. Questions and conversations: 25 multiple-choice items[1] e. Not provided f. A scale score ranging from 0–80[2]	a. Objectivity: standardized input with audio recordings included for administration. Scoring instructions: punched by stencil or the institution's answer sheets[2] b. Reported reliability coefficient (Cronbach's Alpha) of the listening section was .98 (internal consistency) with SEM of 1.91.[5] As part of the development process for Forms D, E, F, equating using anchor items from Form A was conducted.[1] c. Validity evidence: The test was piloted with 480 test takers from 29 different first languages. [1] Test takers (n=312) were distributed throughout the five CEFR levels according to the cut scores in EPT.[5] The Paper-Pencil and Computer Based versions of CaMLA EPT were found to be comparable, regardless of delivery method. [6]	Scores mapped to the CEFR.[2,5] The overall scores of the test providing a skill level: beginner (0-26), beginner high (27-40), intermediate low (41-50), intermediate (51-61), advanced low (62-68), and advanced (69-80)[2]

7. Certificate of Proficiency in English (CPE) a. Cambridge English Language Assessment b. 1913, updated in 1975, 1984, 2002, and 2013[1] c.http://www.cambridgeenglish.org/exams/proficiency/d. Approximately £293 Great Britain Pounds depending on the test center's location[2]	Proficiency test. It is used and accepted as a qualification in near-native English by thousands of leading businesses and educational institutions around the world.[3] The ability to identify gist, detail, function, purpose, topic, speaker, feeling, attitude, opinion, inference, agreement, and main points through a range of spoken material, including conversations, lectures, seminars, broadcasts and talks[1]	Paper-based and computer-based[1] Approximately 40 minutes[1] Heard twice[1] Part 1 (4 short unrelated recordings of monologues and texts with interacting speakers): 6 three-option multiple-choice items; Part 2 (a monologue of an informative nature, aimed at a non-specialist audience): 9 sentence-completion; Part 3 (opinions and attitudes in conversation recording): 5 four-option multiple-choice items; Part 4 (5 short themed monologues): 10 multiple-matching items[4] e. A variety of native speaker accents including British, American, and Australian[1] f. One mark for each correct answer[1]	a. Objectivity: standardized instructions - the instructions are pre-recorded and played automatically. Scoring instructions: machine scoring or scores are given by trained raters depending on the delivery mode.[5] b. Reported reliability coefficients (Cronbach's Alpha) were .92 (internal consistency) with SEM of 2.88 for the total score and .74 (internal consistency) with SEM of 2.18 of the listening section.[6] c. Validity evidence: construct validity, cognitive- and context-related aspects of validity[7]	Targeted at CEFR level C2 (a scale score of 200-230) Grade A (scale score of 220-230), B (scale score of 213-219), and C (scale score of 200-212)[8] Receiving a Statement of Results and a Certificate of Proficiency in English with scale score of 200-230[8] Receiving a Cambridge English certificate demonstrating the ability at CEFR Level C1 with a scale score between 180 and 199[8]
8. Comprehensive Adult Student Assessment System (CASAS) - Life and work listening 980 series a. CASAS b. 1981[1] c. http://www.casas.org/ d. Total package costs US$395.[2]	Achievement test. It is used to measure the basic skills and the English language and literacy skills needed to function effectively at work and in life.[3] The abilities include: phonology, vocabulary, grammar; general discourse, informational discourse, strategies, and critical thinking.[4]	Paper-based or computer-based (online or CD-ROM)[2] Approximately 50 minutes[2] Level A: heard twice; Level B: heard twice in the first two parts, heard once in the last part; Level C: heard once[2] Level A (predicting next line with picture prompt): 38 three-multiple-option items; Levels B and C (predicting next line and identifying true statements): 38 three-multiple-option items[5] Not provided CASAS eTests, TOPSpro Software, or Manual Scoring[5]; a scale score ranging from 162 to 228[5]	a. Objectivity: standardized instructions - the instructions are pre-recorded and played automatically. Scoring instructions: machine scoring or scores are given by proctors depending on the delivery mode.[5] b. Not provided c. Validity evidence: CASAS standardized tests meet the requirements of the Workforce Investment Act (WIA) and correlate with the definitions used in the National Reporting System (NRS) for adult education for ESL programs.[5]	Two series: the 80 series and the 980 series; only the 980 series approved through June 30, 2016 for Measuring Educational Gain in the NRS[6] Scale Score Ranges as beginning ESL (scale score of 162-180), low beginning ESL (scale scores of 181-189), high beginning ESL (scale scores of 190-199), low intermediate ESL (scale scores of 200-209), high intermediate ESL (scale scores of 210-218), advanced ESL (scale scores of 219-227), and exit from advanced ESL (scale scores of 228 and above)[5]

(*Continued*)

(Continued)

TEST NAME	*PURPOSES & LISTENING CONSTRUCTS*	*CHARACTERISTICS OF LISTENING TASKS*	*VALIDITY EVIDENCE*	*COMMENTS*
9. Comprehensive English Language Test (CELT) a. Publisher: Mcgraw-Hill College[1] b. 1986[1] c. http://www.canadiantestcentre.com/OTHER-PRODUCTS/CELT.php d. US$157 for 10 packs of the test books.[1] The test is out of print from the publisher; however, the test can be ordered from the Canadian Test Centre's website[2] and other e-commerce companies.	Placement test. It is used for placement purposes or as entry and exit tests; grade ranging from High School through Adult.[2] The ability to comprehend spoken discourse through short statements, questions, and dialogues: lexical meanings, grammatical meaning where the relationship between form and meaning is relatively direct, both an elaborated exchange and the questions, grammatical meaning with special emphasis on time referencing[3]	Paper-based; Forms A and B[2] Approximately 40 minutes[2] Heard once[1] Task 1 (understanding Wh- or yes/no questions); Task 2 (understanding a sentence involving conditions, comparisons, and time and number expressions); Task 3 (understanding a more elaborated two-turn exchange); 50 four-option multiple-choice items[3] e. Not provided f. Reported scores as the percentage of correct answers[1]	a. Objectivity: standardized instructions - Not provided. Scoring instructions: machine scoring or scores are given by proctors.[1] b. Reported reliability coefficient was .98 (internal consistency) with SEM of 7.05 for the total score.[5] KR-20 reliability estimates for the overall test scores ranged from .88 to .96, with SEM of 3.85.[3] c. Validity evidence: There was a correlation of .79 between the TOEFL and the CELT and .81 with the Michigan Test of English Language Proficiency.[1] Slark and Bateman (1982, as cited in Graham, 1987) used class grades as the criterion for academic success and found that CELT Listening scores correlated significantly with grades in 9 out of 22 courses.[5]	Lacks authenticity;[1] however, can be used to determine students' listening proficiency (e.g., In'nami, 2006[6]) Low-to-moderate stakes settings[1] Scores reported in terms of listening proficiency levels: 20-35 (High-intermediate listening proficient) and 8 - 18 (Low-intermediate listening proficient)[7]

10. Computerized Adaptive Placement Exams (CAPE, or WebCAPE) a. First developed by Brigham Young University and has been licensed WebCAPE to Perpetual Technology Group[1] b. 1999[2] c. http://www.perpetualworks.com/webcape/overview d. Educators can try the exams free for 30 days. For three ESL exams: Grammar, Reading, and Listening, one time license and set up fee cost US$500.[3]	Placement and Proficiency test. It is used for academic (placement test) or business (employees' proficiency levels) goals.[4] The ability to understand the main idea, identify specific details, and draw inferences based on the whole passage.[2]	Computer-adaptive[5] Varying time but approximately 20–25 minutes[4] c. Heard once d. Listening to a fairly short passages; level checkers in the first six items; finish when the test takers incorrectly answering four items at the same difficulty level, or answering five items at the highest difficulty level possible[4] e. Not provided f. Automatic scoring; scores reported immediately after the test[4]	a. Objectivity: standardized instructions - the instructions are pre-recorded and played automatically.[5] Scoring instructions: machine scoring automatically.[5] b. Reported reliability coefficients of the overall test ranged from .76–.86 (test-retest).[5] c. Validity evidence: WebCAPE English Language Assessment has been calibrated in accordance with the standards American Council for the Teaching of Foreign Languages (ACTFL) proficiency guidelines.[5]	The computer displaying the performance level, corresponding with English courses (the first two years of college language courses) or depending on the designated institution[5]
11. EF Standard English Test (EFSET) a. EF Education First[2] b. 2013[5] c. https://www.efset.org/en d. Free[2]	a. Proficiency test. It is used as a free online standardized test of the English language designed for non-native English speakers.[1] Assess receptive skills only (reading and listening comprehension) and do not assess writing or speaking.[3] b. The ability is understanding conversation between native speakers.[5]	a. Computer-adaptive[5] b. Varying time but approximately 25–60 minutes[2] c. Heard once[2] d. Multiple right answers, matching, and categorization; number of items depending on students' language proficiency[5] e. A variety of accents[5] f. Scores ranging from 0 to 100, with CEFR, IELTS and TOEFL equivalencies given[5]	a. Objectivity: standardized instructions - the instructions are pre-recorded and played automatically.[5] Scoring instructions: machine scoring automatically.[5] b. Not provided c. Validity evidence: The EFSET is a standardized objectively-scored test of listening and reading skills. It is designed to classify test takers' reading and listening performances on the test into one of the 6 levels established by the CEFR.[5] EF began the design process by soliciting the assistance of language assessment experts, and engaging in a formal, highly structured design process.[5]	Two versions of the EFSET: a 50 minute test assigning a score on the 6-level Common European Framework of Reference for Languages ("EFSET"), and a 2 hour test assigning a score from 0 to 100 with a TOEFL and IELTS equivalency score in addition to the 6-level CEFR score ("EFSET PLUS")[4]

(*Continued*)

(Continued)

TEST NAME	PURPOSES & LISTENING CONSTRUCTS	CHARACTERISTICS OF LISTENING TASKS	VALIDITY EVIDENCE	COMMENTS
12. First Certificate in English (FCE) a. Cambridge English Language Assessment[1] b. 1939; regular updates[2] c. http://www.cambridgeenglish.org/exams/first/ d. Approximately £120-150 British Pounds depending on location[3]	a. Proficiency test. It is used for learners who want to: start working in an English-speaking environment; study at an upper intermediate level, such as foundation or pathway courses and live independently in an English-speaking country.[2] b. The ability to listen to a range of spoken material, including lectures, radio broadcasts, speeches and talks, and understand feeling, attitude, detail, opinion, purpose, agreement, gist, function, topic, specific information, etc.[2]	Paper-based or computer-based[2] b. Approximately 40 minutes[2] c. Heard twice[2] d. Part 1 (8 short unrelated extracts from monologues or dialogues): 1 multiple-choice item for each; Part 2 (a monologue): 10 items with a sentence-completion task; Part 3 (5 short related texts): 8 multiple-matching options; Part 4 (an interview between two speakers): 7 three-option multiple-choice items[2] e. A variety of native speaker accents including British, American, and Australian[2] f. Exact match scoring (1 point for each correct answer)[2]; the Statement of Results containing scores on Cambridge English Scale (from 80–230) for each section as well as overall score[2]	a. Objectivity: standardized instruction - the instructions are scripted and given by trained proctors OR the instructions are pre-recorded and played automatically depending on the delivery mode.[2] Scoring instructions: machine scoring OR scores are given by trained raters depending on the delivery mode.[1] b. Reported reliability coefficients (Cronbach's Alpha) were .94 (internal consistency) with SEM of 2.78 for the total score and .81 (internal consistency) with SEM of 2.16 of the listening section.[6] c. Validity evidence: All Cambridge systems and processes for designing, developing and delivering exams and assessment services are certified as meeting the internationally recognized standard for quality management.[5]	Targeted at Level B1 – the second highest level on the CEFR scale[2] Results available online for free and verified by institutions instantly[4] Receiving a certificate with performance ranging between CEFR Levels B1 and C2[2]

13. International English Language Testing System (IELTS) a. Jointly owned by British Council, IDP, IELTS Australia, and Cambridge English Language Assessment b. 1989[1] c. http://www.ielts.org/ d. Approximately £115 British Pounds, €190 euros, or US$200, depending on location[2]	a. Proficiency test. It is used as a proficiency test of English that has two versions: General and Academic (Reading and Writing section have tasks related to academic topics). It assesses the four skills: Listening, Reading, Writing, and Speaking.[3,4] b. The ability to understand main ideas and detailed factual information, ability to understand the opinions and attitudes of speakers, ability to understand the purpose of an utterance, and ability to follow the development of ideas.[4]	a. Paper-based and computer-based b. Approximately 30 minutes c. Heard once d. Four sections with monologues and dialogues in various social and academic contexts: each followed by 10 items of different types: Multiple choice, matching, plan/map/diagram labeling, form/note/table/flow-chart/summary completion, sentence completion; Part 1(conversation about social context), Part 2 (monologue about social context), Part 3 (conversation on academic subjects), Part 4 (monologue on academic subjects)[9] e. A variety of native speaker accents including British, American, Australian, and New Zealand[9] f. Reported score by bands from (1 nonuser to 9 expert user); a score of 0 for no assessable information provided[4]	a. Objectivity: standardized instructions and listening input through recordings. Scoring instructions: answers are marked by certified markers at test center. Then answer sheets are sent to Cambridge English Assessment for analysis.[3] b. Reported reliability coefficients (Cronbach's Alpha) was .90 (internal consistency) with SEM of .39 of the listening section.[5] c. Validity evidence: Predictive validity of academic performance: small to medium effect.[6,7] There are numerous studies on validity, candidate performance, stakeholder attitudes, and other related topics using IELTS data.[8]	More than 1,000 locations in 140 countries

(*Continued*)

(Continued)

TEST NAME	*PURPOSES & LISTENING CONSTRUCTS*	*CHARACTERISTICS OF LISTENING TASKS*	*VALIDITY EVIDENCE*	*COMMENTS*
14. Michigan English Language Assessment Battery (MELAB) a. Cambridge Michigan Language Assessments b. 1985[1] c. http://www.cambridgemichigan.org/institutions/products-services/tests/proficiency-certification/melab/[2] d. Depending on location.[1] Approximately US$75–95.	a. Proficiency test. It is used to evaluate English language proficiency at an advanced level for academic or professional purposes.[2] b. The ability to master three types of questions: global, local, and inferential: Global-Understanding main idea, identifying speaker's purpose, synthesizing ideas from different parts of the stimulus; Local-Identifying supporting detail, understanding vocabulary, synthesizing details, recognizing restatement; Inferential-Understanding rhetorical function, making an inference, inferring, supporting detail, understanding pragmatic implications.[2]	a. Paper-based[2] b. Approximately 30-35 minutes[2] c. Heard once[2] d. Part 1 (a short recorded question or statement): 18 three-option multiple-choice items; Part 2 (A recorded conversation): 22 three-option multiple-choice items; Part 3 (Four recorded interviews): 20 multiple-choice items[2] e. Not provided f. Each correct answer contributing proportionally to each section then scaled on a range from 0–100[2]	a. Objectivity: standardized input in audio recorded input administered at test centers. Scoring instructions: computer scored[3] b. Reported reliability of the listening section was .88 (internal consistency) using KR-20, and SEM estimate was 4.53.[4] c. Validity evidence: The MELAB assesses mastery of skills at the B1 to C1 CEFR levels.[2] MELAB and TOEFL CBT scores have a significant correlation of .89.[5] Listening construct validity of MELAB was confirmed through a Confirmatory Factor Analysis (CFA).[8]	Valid scores for two years[2]

15. Michigan Test of English Language Proficiency Series (MTELPs Series) a. Cambridge Michigan Language Assessments b. 2014[1] c. http://www.cambridgemichigan.org/institutions/products-services/tests/placement-progress/MTELP-series/ d. The cost of the whole package (180 booklets) is US$3,945, and the cost of 25 computer-based test is US$198.75.[2]	a. Achievement test. It is used to monitor progress and assess achievement. This series has three levels. Based on scores (and institution's policies and courses), the test takers may be classified (or placed) into finding the specific level too difficult, appropriate, or too easy. To decide which level of the test to administer, the institution can look at scores in CaMLA EPT, previous courses, or previous achievement.[2,3] It assesses listening, grammar, vocabulary, and reading at the beginner, intermediate and advanced levels separately. [2] b. The abilities include understanding main idea, inferencing, and details.[4,5,6]	a. Paper-based and computer-based[2] b. Approximately 50 minutes[2] c. Heard once[4, 5 6] d. Level 1 (a short conversation and a question or a statement): 20 three-option multiple-choice items; Level 2 (a short conversation between 2 speakers and a short talk): 25 three-option multiple-choice items; Level 3 (a conversation between 2 speakers and an interview featuring several speakers): 25 three-option multiple-choice items[6] e. Not provided f. The overall test scores ranging from 0–100[2]	a. Objectivity: test directions are scripted for test administrators in the administration manual. Scoring instructions: tests are scored by institutions who purchase tests using stencils and answer sheets[2] b. Not provided c. Validity evidence: The scores are mapped to the CEFR levels but no study has been done to corroborate the CEFR cut scores suggested.[3]	Parallel forms available within each level[2] The scaled scores ranging from Level 1, appropriate (2-49), too easy (50-100). Level 2, too difficult (6-25), appropriate (25-74), too easy (75-100). Level 3, too difficult (15-50), appropriate (51-100)[3]

(*Continued*)

(Continued)

TEST NAME	PURPOSES & LISTENING CONSTRUCTS	CHARACTERISTICS OF LISTENING TASKS	VALIDITY EVIDENCE	COMMENTS
16. Pearson Test of English Academic (PTEA) a. Pearson Company b. 2009[7] c. http://pearsonpte.com/Pages/Home.aspx d. Approximately US$200, depending, on location[6]	a. Proficiency test. It is used to accurately measure the listening, reading, speaking, and writing skills of test takers who are non-native speakers of English and need to demonstrate their level of academic English proficiency. It delivers a real-life measure of test takers' language ability to universities, higher education institutions, government departments and organizations requiring academic English. It is a new, international, computer-based academic English language test.[1] b. The ability to understand both explicit and implicit information, and both concrete and abstract information (e.g., identify the purpose, understand the main ideas, detect the tone and attitude, identify text structure, understand the communicative function, infer the conceptual framework and relationships within discourses, extract salient points, draw valid inferences and conclusions, and integrate information from multiple sources).[8]	a. Computer-based[1] b. Approximately 45–57 minutes[1] c. Heard once[3] d. 17–25 questions[1] Summarize spoken text (Listening & Writing, 2–3 items); multiple-choice-single answer (Listening, 2–3 items); multiple-choice-multiple answers (Listening, 2–3 items); fill in the blanks (Listening & Writing, 2–3 items); highlight correct summary (Listening & Reading, 2–3 items); select missing word (Listening, 2–3 items); highlight incorrect words (Listening & Reading, 2–3 items); write from dictation (Listening & Writing, 3–4 items)[2] Can adjust the volume on each item[3] e. A variety of native speaker accents including British, American, and Australian[3] f. Computer-scoring; scores ranging from 10–90; dichotomous and partial credits[2]	Objectivity: standardized instruction - the instructions are pre-recorded and played automatically.[2] Scoring instructions: machine scoring, and scores are given by two or three independent trained raters.[2] Reported reliability correlations of the overall test was .92 (split-half) of the overall score and .89 (split-half) of the listening section.[5] Validity evidence: The test was linked to other external frameworks of language proficiency so PTE was benchmarked to the CEFR.[5] Topics were selected to cover a wide range of academic contexts but avoiding texts that require specific domain knowledge that may cause bias toward certain test takers.[5] All test items and tasks underwent a series of review processes at multiple stages, including author and peer review, sensitivity analysis, internal review, item migration quality checks and item pool review.[5]	Successful undergraduate studies: 51-61; successful postgraduate studies:57-67; successful MBA studies:59-69[3]

17. Test of Adult Basic Education Complete Language Assessment System–English (TABE CLAS-E) a. McGraw-Hill Education CTB b. 2007[1] c. http://www.ctb.com/ctb.com/control/ctbProductViewAction?p=products&productId=865 d. The test booklets with CD cost US$101.50 per 25 units.[2]	Proficiency test. It is used to measure adult learners' English language proficiency and aid in transitioning learners into mainstream education programs or career paths.[3] It can also be used for placement purposes.[4] The abilities include Quantitative Literacy (Numbers and Numeracy Terms), Listen for Information (Discern Sounds, Details, Stated Concepts), Interpersonal Skills (Idiom/Expression, Determine Roles, Instructions), Interpret Meaning (Cause/Effect, Fact/Opinion, Main Idea, Forecast, Speaker Purpose).[3]	Paper-based[3] Approximately 20 minutes[3] c. Not provided d. A spoken format of varying lengths in the context of meaningful work, community, and education situations; 25 three-option multiple-choice items[5] e. Not provided f. One mark for each correct answer; hand scoring or computer-based scoring; score reports including Individual Student Report, Item Analysis Report, Pretest and Posttest Report, Prescriptive Report, Group List Report, Assessment Summary Report, and Rank List Report[3]	a. Objectivity: standardized instruction - the instructions are pre-recorded and played automatically.[3] Scoring instructions: machine scoring or scores are given by proctors.[3] b. Not provided c. Validity evidence: The test items are aligned with key U.S. standards for adult education ESL, as well as the CEFR and the National Reporting System (NRS) core measure.[3]	4 assessment levels: Level 1 (Beginning ESL 1), Level 2 (Beginning ESL 2), Level 3 (Intermediate ESL), and Level 4 (Advanced ESL)[5]

(*Continued*)

(Continued)

TEST NAME	PURPOSES & LISTENING CONSTRUCTS	CHARACTERISTICS OF LISTENING TASKS	VALIDITY EVIDENCE	COMMENTS
18. Test of English as a Foreign Language-Internet Based Test (TOEFL iBT) a. Educational Testing Service b. September 2005[1] c. http://www.ets.org/toefl d. Approximately US$180, depending on location[8]	a. Proficiency test. It is used to provide evidence for the English language proficiency of non-native English speaker test takers to the higher learning institutions with English as their language of instruction. Scores are also used by government agencies, scholarship and internship programs.[2] b. The abilities include understanding main idea, detail, speakers' attitude or function, organization, relationships between the ideas and inference.[3]	a. Computer-based[3] b. Approximately 60–90 minutes[3] c. Heard once d. 34–51 questions[3] 4–6 lectures (3–5 minutes for each) 2–3 conversations (3 minutes for each)[3] four question formats: 4-option multiple-choice items with a single correct answer; multiple-choice items with more than one correct answer; ordering events or steps in a process; matching objects or text to categories in a chart[3] e. A variety of native-speaker accents including British, North American, Australian, and New Zealand[9] f. Scores ranging from 0–30; dichotomous and partial scoring[4]	Objectivity: standardized instruction - the instructions are pre-recorded and played automatically.[3] Scoring instructions: scores are given by machine and trained raters.[3] Reported reliabilities were .94 (generalizability coefficient) with SEM of 5.64 for the overall score and .85 (generalizability coefficient) with SEM of 3.20 for the listening section.[5] Validity evidence: This evidence is collected through studies on test content, scoring processes, relationships to other measures of proficiency and the impact on teaching and learning English.[6] The test development process may take from 6 to 18 months, to ensure that tests and items meet strict quality standards and that test forms are similar to each other in content and difficulty.[3] The content of the test is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings.[7]	Low (0-14); Intermediate (15-21); High (22-30)

19. Test of English for International Communication (TOEIC) Listening & Reading Test a. Educational Testing Service b. 2006[2] c. http://www.etsglobal.org/Tests-Preparation/The-TOEIC-Tests d. Approximately US$75, depending on location[1]	a. Proficiency test. It is used to assess the ability to use English in real-life work situations. In addition, the test design ensures that scores can be accurately compared between candidates around the globe.[1] b. The abilities to understand explanations and instructions, understand workplace conversations, and follow public talks and announcements.[1]	a. Paper-based and computer-based[3] b. Approximately 45 minutes[1] c. Heard once[1] d. 100 questions Part 1: photographs; Part 2: question-response; Part 3: conversations; Part 4: short talks[1] e. A variety of accents including British, American, Canadian, and Australian[3] f. Scores ranging from 5–495 points[1]	a. Objectivity: standardized instruction - the instructions are scripted and given by trained proctors in the paper-based version.[1] Scoring instructions: machine scoring.[1] b. Reported reliability coefficients (Cronbach's Alpha) were .95–.96 (internal consistency) with SEM of 25 for the total score and .92–.93 (internal consistency) of the listening section.[2] c. Validity evidence: Evidence that the TOEIC measures English-language proficiency comes first from the careful way in which language testing experts design and assemble the test so as to include a variety of important English-language tasks.[1]	If students take several versions of the test within a short period of time, they will obtain a number of scores that center around an average value known as "true" score.[1] Because of limited space in the score report, only statements of strength are provided in the report.[1] Listening and reading scores are added together and correlated to the CEFR levels A1 to C: 945 points-C1, 785 points-B2, 550 points-B1, 225 points-A2, 120 points-A1.[4]
20. Woodcock Language Proficiency Battery-Revised (WLPB-R) a. Riverside Publishing Company[1] b. 1991[1] c. http://ericae.net/eac/eac0187.htm d. Purchase a complete program: US $1,149.99[1]	a. Proficiency test. It is used as an overall measure of language proficiency and greatly expanded measures of oral language, reading, and written language in both English and Spanish. The WLPB-R English Form and Spanish Form are parallel versions which facilitates comparison between the languages.[2] b. The ability to comprehend a passage and supply the single word missing at the end in an oral cloze procedure.[3]	a. Computer-based[1] b. Approximately 20–60 minutes[2] c. Not provided d. The listening section beginning with simple verbal analogies and associations and progressing to a higher level of comprehension involving the ability to discern implications[3] e. Not provided f. Available scores including: age and grade equivalents, standard scores, percentile ranks, Relative Proficiency Indexes (RPI), instructional ranges and a Comparative Language Index (CLI) when both languages are being administered[2]	a. Objectivity: standardized instruction—the instructions are prerecorded and played automatically.[2] Scoring instructions: machine scoring[2] b. Cluster reliabilities are in the .90s for the overall test.[2] c. Validity evidence: The Examiner's Manual documents the validity of the WLPB-R as a technically sound measure of language proficiency.[2]	The WLPB-R composed of 13 tests in three domains: 5 of oral language, 4 of reading, and 4 of written language[2] Available scores include: age and grade equivalents, standard scores, percentile ranks, Relative Proficiency Indexes (RPI), instructional ranges, CALP levels (Spanish only), and a Comparative Language Index (CLI) when both languages have been administered.[2]

References per Test

1. ACT ESL Placement Test

ACT Compass. English as a Second Language (ESL) Placement Test. [Website] (2015). Retrieved from http://www.act.org/compass/tests/esl.html

Stoynoff, S., & Chapelle, C. A. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.

ACT Compass. Internet Version Reference Manual [PDF]. (2012) Retrieved from http://www.act.org/compass/pdf/CompassReferenceManual.pdf

The ACT Test. User Handbook for Educators [PDF]. (2015). Retrieved from http://www.act.org/content/dam/act/unsecured/documents/ACT-UserHandbook.pdf

2. Business Language Testing Service (BULATS)
BULATS. Business Testing Language Service. [Website] (2015). Retrieved form http://www.bulats.org/
Stoynoff, S., & Chapelle, C. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.
BULATS. Why BULATS? [Website] (2015). Retrieved from http://www.bulats.org/why-bulats
BULATS. Information for candidates. [Website] (2011). Retrieved from http://www.bulats.org/sites/bulats.org/files/info_cand_en.pdf
Clark, D. (2006). Essential BULATS. *The Bulats listening test* (pp. 5–23). [PDF excerpt]. Retrieved from www.cambridge.org/download_file/697596/0/
Cope, L. (2009). CB BULATS: Examining the reliability of computer-based test. *Cambridge ESOL: Research Notes, 38*, 31-34.
Principles of good practice: Quality management and validation in language assessment [PDF]. (2013). Retrieved from http://www.cambridgeenglish.org/images/22695-principles-of-good-practice.pdf
3. Cambridge English: Advanced (CAE)
Cambridge English: Advanced (CAE). [Website] (2015). Retrieved from http://www.cambridgeenglish.org/exams/advanced/
Cambridge English Advanced. Handbook for Teachers for Exams from 2015 [PDF]. Retrieved from http://www.cambridgeenglish.org/images/168194-cambridge-english-proficiency-teachers-handbook.pdf
FAQs about Cambridge English: Advanced (CAE). [Website]. (2015). Retrieved from http://www.cambridgeenglish.org/exams/advanced/faqs/#A2.5
Results and certificates for Cambridge English: Advanced (CAE) [Website]. (2015). Retrieved from http://www.cambridgeenglish.org/exams/advanced/results/
Principles of good practice: Quality management and validation in language assessment [PDF]. (2013). Retrieved from http://www.cambridgeenglish.org/images/22695-principles-of-good-practice.pdf
Quality and accountability [Website]. (2015). Retrieved from http://www.cambridgeenglish.org/principles/
4. Canadian Academic English Language Assessment (CAEL)
The Canadian Academic English Language Assessment. What is CAEL Assessment? [Website] (2015). Retrieved from https://www.cael.ca/edu/whatis.shtml
Stoynoff, S., & Chapelle, C. A. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.
The Canadian Academic English Language Assessment. Test Score and User's Guide [PDF]. (2015). Retrieved from https://www.cael.ca/wp-content/uploads/2016/01/Users-Guide-Paragon-November-2015.pdf
5. Cambridge Michigan Examination for the Certificate of Proficiency in English (ECPE)
English as a Second Language Directory [Website]. (2015). Retrieved from http://www.esldirectory.com/blog/english-language-exams/ecpe/
Cambridge Michigan ECPE [Website]. (2014). Retrieved from http://examenglish.com/ECPE/index.php
ECPE [Website]. (n.d.). Retrieved from http://www.cambridgemichigan.org/institutions/products-services/tests/proficiency-certification/ecpe/
Cambridge Michigan Language Assessments. (2013). *2009-2010 Technical Review* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/ECPE_TechReview_2009-2010.pdf
Cambridge Michigan Language Assessments. (2015). *ECPE 2014 Report* [PDF]. Retrieved from https://www.cambridgemichigan.org/wp-content/uploads/2015/04/ECPE-2014-Report.pdf
Lin, C. (2015). A practical guide to investigating score reliability under a Generalizability Theory framework. *CaMLA Working Papers, 15*, 1-11.
6. Cambridge Michigan English Placement Test (CaMLA EPT)
Walter, D., & Hentschel, J. (2013). *CaMLA English Placement Test (EPT) forms D-F: Development report* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/EPT-Development-Report-20131.pdf
CaMLA EPT [Website]. (n.d.). Retrieved from http://www.cambridgemichigan.org/institutions/products-services/tests/placement-progress/camla-ept/
Cambridge Michigan Language Assessments. (2013). *EPT Sample Exam Items* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/EPT_SampleItemBooklet_2013.pdf
Cambridge Michigan Language Assessments. (2013). *EPT Quality in Test Design* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/Flyer-EPT-Quality-in-Test-Design.pdf

Cambridge Michigan Language Assessments. (2013). *Linking the Common European Framework of Reference and the CaMLA English Placement Test: Technical Report* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/EPT-Technical-Report-20140625.pdf

Cambridge Michigan Language Assessments. (2014). *Comparing the Paper-Based and Computer-Based CaMLA EPT* [PDF].Retrieved from https://www.cambridgemichigan.org/wp-content/uploads/2014/12/ept-pb-cb-comparison-20141222.pdf

7. Certificate of Proficiency in English (CPE)

Cambridge English Proficiency: Handbook for teachers [PDF]. (2015). Retrieved from http://www.cambridgeenglish.org/images/168194-cambridge-english-proficiency-teachers-handbook.pdf

Upcoming examinations [Website]. (2015). Retrieved from http://www.cambridgeopencentre.org/asp_pages/examinations.asp

Cambridge English: Proficiency (CPE) [Website] (2015). Retrieved from http://www.cambridgeenglish.org/exams/proficiency/

Exam format [Website] (2015). Retrieved from http://www.cambridgeenglish.org/exams/proficiency/exam-format/

Preparation [Website] (2016). Retrieved from http://www.cambridgeenglish.org/exams/proficiency/preparation/

Quality and accountability [Website]. (2015). Retrieved from http://www.cambridgeenglish.org/principles/

Principles of good practice: Quality management and validation in language assessment [PDF]. (2013). Retrieved from http://www.cambridgeenglish.org/images/22695-principles-of-good-practice.pdf

Results [Website] (2015). Retrieved from http://www.cambridgeenglish.org/exams/proficiency/results/

8. Comprehensive Adult Student Assessment System (CASAS)

CASAS history [Website]. (2015). Retrieved from https://www.casas.org/about-casas/history

CASAS 2016 [Catalog]. (2016). Retrieved from https://www.casas.org/docs/newsroom/Catalog.pdf?Status=Master

What is CASAS? [Brochure] (2015). Retrieved from https://www.casas.org/docs/default-source/pagecontents/download-about-casas.pdf?sfvrsn=7

CASAS Listening Basic Skills Content Standards [Supplemental material]. (2009). Retrieved from https://www.casas.org/docs/research/listening-content-standards.pdf?sfvrsn=3?Status=Master

Life and work listening [Website]. (2015). Retrieved from https://www.casas.org/product-overviews/assessments/life-and-work-listening

Using CASAS to meet NRS accountability requirements [Website]. (2015). Retrieved from https://www.casas.org/training-and-support/wia-and-nrs-compliance

Life and work listening 980 series [Supplemental material]. (2015). Retrieved from https://www.casas.org/docs/default-source/pagecontents/life-and-work-listening-980-series-letter-to-field.pdf?sfvrsn=20?Status=Master

9. Comprehensive English Language Test (CELT)

Stoynoff, S., & Chapelle, C. A. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.

CELT [Website]. (2015). Retrieved from http://www.canadiantestcentre.com/OTHER-PRODUCTS/CELT-TestBooks.php

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

Mills, A., Swain. L., & Weschler, R. (1996). The implementation of a first year English placement system. *The Internet TESL Journal*, 2. Retrieved from http://iteslj.org/Articles/Mills-Placement.html

Graham, J. G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly, 21*, 505–521.

In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System, 34*, 317-340. doi:http://dx.doi.org.libproxy.nau.edu/10.1016/j.system.2006.04.005

Chiang, C. S., & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly, 26*, 345–374.

10. Computerized Adaptive Placement Exams (CAPE, or WebCAPE)

WebCAPE language placement exams [Website]. (n.d.). Retrieved from http://www.perpetualworks.com/about

Stoynoff, S., & Chapelle, C. A. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.

Pricing [Website]. (n.d.). Retrieved from http://www.perpetualworks.com/webcape/pricing

WebCAPE's ESL placement test [Website]. (n.d.). Retrieved from http://www.perpetualworks.com/languages/esl/

WebCAPE's details [Website]. (n.d.). Retrieved from http://www.perpetualworks.com/webcape/details

11. EF Standard English Test (EFSET)

Take a new test aimed at the world's English language learners. (2014, October 1). *National Public Radio*. Retrieved from http://www.npr.org/blogs/goatsandsoda/2014/10/01/352983784/take-a-new-test-aimed-at-the-worlds-english-language-learners

EF Standard English Test [Website]. (*n.d.*). Retrieved from https://www.efset.org/en

Testing times for English teaching firm as it taps examination market. (2014, October 9). *The China Daily*. Retrieved from http://www.ecns.cn/business/2014/10-09/137542.shtml

Free online test targets English learners. (2014, October 17). *Voice of America*. Retrieved from http://m.learningenglish.voanews.com/a/efset-launch-free-online-test/2485916.html

EF SET. (2014). *EFSET technical background report* [PDF]. Retrieved from http://cdn.efset.org/media/EFSET_Technical_Background_Report_2014.pdf

12. Cambridge English: First (FCE)

Cambridge English: First (FCE) [Website]. (2015). Retrieved from http://www.cambridgeenglish.org/exams/first/

Cambridge English: First. Handbook for Teachers for Exams from 2015[PDF]. (2015). Retrieved from http://www.cambridgeenglish.org/images/cambridge-english-first-handbook-2015.pdf

How to Register for Cambridge English: First (FCE) [Website]. (2015). Retrieved from http://www.cambridgeenglish.org/exams/first/how-to-register/

Results and certificates for Cambridge English: First (FCE) [Website]. (2015). Retrieved from http://www.cambridgeenglish.org/exams/first/results/

Principles of good practice: Quality management and validation in language assessment [PDF]. (2013). Retrieved from http://www.cambridgeenglish.org/images/22695-principles-of-good-practice.pdf

Quality and accountability [Website]. (2015). Retrieved from http://www.cambridgeenglish.org/principles/

13. International English Language Testing System (IELTS)

History of IELTS [Website]. (n.d.). Retrieved from http://www.ielts.org/researchers/history_of_ielts.aspx

Exam English [Website]. (2014). Retrieved from http://www.examenglish.com/IELTS/

IELTS [Website]. (n.d.). Retrieved from http://www.ielts.org/test_takers_information/what_is_ielts.aspx

IELTS. (2013). *Guide for educational institutions, governments, professional bodies and commercial organizations* [PDF]. Retrieved from http://www.ielts.org/pdf/guide_edu-%20inst_gov_2013.pdf

IELTS [Website]. (n.d.). Retrieved from http://www.ielts.org/researchers/analysis_of_test_data/test_performance_2013.aspx

Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a group of international students at the University of Tasmania. *IELTS Research Reports, 1*, 72–115.

Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. *IELTS Research Reports, 3*, 85–108.

IELTS [Website]. (*n. d.*). Retrieved from http://www.ielts.org/researchers/research.aspx

Cambridge English Language Assessment [Website]. (n.d.). Retrieved from https://www.teachers.cambridgeesol.org/ts/exams/academicandprofessional/ielts

14. Michigan English Language Assessment Battery (MELAB)

Stoynoff, S., & Chapelle, C. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.

MELAB [Website]. (n.d.). Retrieved from http://www.cambridgemichigan.org/institutions/products-services/tests/proficiency-certification/melab/

MELAB [Website]. (n.d.). Retrieved from http://www.cambridgemichigan.org/test-takers/tests/melab/

Cambridge Michigan Language Assessments. (2015). *MELAB 2014 Report* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2015/03/MELAB-2014-Report.pdf

Dobson, B., Han, I., & Yamashiro, A. D. (2001). *MELAB/Computer-Based TOEFL Study* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/MELAB_TOEFLStudy_2001.pdf

Wang, S. (2006). Validation and invariance of factor structure of the ECPE and MELAB across gender. *SPAAN FELLOW Working Papers in Second or Foreign Language Assessment, 4*, 41–56.

Eom, M. (2008). Underlying factors of MELAB listening constructs. *SPAAN FELLOW Working Papers in Second or Foreign Language Assessment, 6*, 77–94.

Goh, C., & Aryadoust, V. (2010). Investigating the construct validity of the MELAB listening test through the Rasch analysis and Correlated Uniqueness Modeling. *SPAAN FELLOW Working Papers in Second or Foreign Language Assessment, 8*, 31–68.

15. Michigan Test of English Language Proficiency Series (MTELPs)

Cambridge Michigan Language Assessments. (2013). *A new suite of English language assessments!* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/MTELPs2013_Announcement.pdf

MTELP Series [Website]. (n.d.). Retrieved from http://www.cambridgemichigan.org/institutions/products-services/tests/placement-progress/mtelp-series/

Using the MTELP series: Levels & Score Interpretation [Website]. Retrieved from http://www.cambridgemichigan.org/institutions/products-services/tests/placement-progress/MTELP-series/levels-scoring/

Cambridge Michigan Language Assessments. (2014). *MTELP Series Level 1 Sample Exam Items* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/10/MTELP-Samples-L1.pdf

Cambridge Michigan Language Assessments. (2014). *MTELP Series Level 2 Sample Exam Items* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/10/MTELP-Samples-L2.pdf

Cambridge Michigan Language Assessments. (2014). *MTELP Series Level 3 Sample Exam Items* [PDF]. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/10/MTELP-Samples-L3.pdf

16. Pearson Test of English Academic (PTEA)

PTE Academic. (2012). *PTE Academic test taker handbook* [PDF]. Retrieved from http://pearsonpte.com/Testme/Documents/PTEA_Test_Taker_Handbook_EN.pdf

PTE Academic. (2012). *Academic score guide* [PDF]. Retrieved from http://pearsonpte.com/PTEAcademic/scores/Documents/PTEA_Score_Guide.pdf

PTE Academic. (2011). *PTE Academic tutorial* [PDF]. Retrieved from http://pearsonpte.com/PTEAcademic/Tutorial/Documents/PTEA_Tutorial.pdf

PTE Academic. (2012). *Using PTE Academic scores* [PDF]. Retrieved from http://www.pearsonpte.com/SiteCollectionDocuments/6747_US_using_PTEAScores_14_10_09_V4.pdf

PTE Academic. (*n.d.*).*Validity and reliability in PTE Academic scores* [PDF]. Retrieved from http://www.pearsonpte.com/SiteCollectionDocuments/ValidityReportUS.pdf

PTE Academic. (2012). *Test of English fees* [PDF]. *PTE Academic*. Retrieved from http://pearsonpte.com/TestMe/Taking/Pages/TestCentersandFees.aspx

Zheng, Y., & Jong, J. (n.d.). Research note: Establishing construct and concurrent validity of Pearson Test of English Academic [PDF]. *Pearson Always Learning*. Retrieved from http://www.pearsonpte.com/research/Documents/PTEAcademicValidityPaper.pdf

PTE Academic. (2012). PTE Academic technical manual [PDF].

17. Test of Adult Basic Education Complete Language Assessment System–English (TABE CLAS–E)

Our heritage [Website]. (2015). Retrieved from http://www.ctb.com/ctb.com/control/ourHeritageAction?p=aboutUs

Browse TABE CLAS-E offerings [Website]. (2015). Retrieved from http://www.ctb.com/ctb.com/control/childNodesViewAction?categoryId=1145&adjBrd=Y

Product QuickFacts [Website]. (2015). Retrieved from http://www.ctb.com/ctb.com/control/ctbProductViewAction?p=products&productId=865#

TABE CLAS-E: FAQs [Website]. (2015). Retrieved from https://www.ctb.com/ctb.com/control/faqAnswerAction?supportCenterId=13721&faqId=13992.0&p=support

Brochures [PDF]. (2015). Retrieved from http://www.ctb.com/ctb.com/control/openFileShowAction?mediaId=869

18. Test of English as a Foreign Language-Internet Based Test (TOEFL iBT)

Contact information [Website]. (2015). Retrieved from http://www.ets.org/toefl/contact/region12

About the test [Website]. (2015). Retrieved from http://www.ets.org/toefl/ibt/about?WT.ac=toeflhome_ibtabout2_121127

TOEFL iBT test framework and test development. (*n.d.). TOEFL iBT Research Insight, 1*. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_research_insight.pdf

Interpret scores [Website]. (2015). Retrieved from http://www.ets.org/toefl/english_programs/scores/interpret/

Reliability and comparability of TOEFL iBT scores [PDF]. (*n.d.). TOEFL iBT Research Insight, 1*. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf

Scores [Website]. (2015). Retrieved from http://www.ets.org/toefl/english_programs/scores/

Validity evidence supporting test score interpretation and use (*n.d.*). *TOEFL iBT Research Insight, 1*. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v4.pdf

Select your TOEFL testing location [Website]. (2015). Retrieved from http://www.ets.org/bin/getprogram.cgi?test=toefl

TOEFL [Website]. (2015). Retrieved from https://www.ets.org/toefl/important_update/english_accents_added

19. Test of English for International Communication (TOEIC) Listening & Reading Test

TOEIC. *(n.d.). TOEIC examinee handbook: listening & reading* [PDF]. Retrieved from https://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf

Liao, C., Hatrak, N., & Yu, F. (2010). Comparison of content, item statistics and test-taker performance for the resigned and classic TOEIC Listening and Speaking Tests [PDF].*TOEIC Compendium Study*. Retrieved from https://www.ets.org/Media/Research/pdf/TC-10-04.pdf

ETS Global: The TOEIC listening and reading test [website]. (2012). Retrieved from http://www.etsglobal.org/Tests-Preparation/The-TOEIC-Tests/TOEIC-Listening-Reading-Test

TOEIC. (2011). *Correlation table: TOEIC listening and reading scores descriptors and the CEFR levels*. Retrieved from https://www.etsglobal.org/content/download/768/12037/version/6/file/TOEIC+L%26R+Descriptors-MAR089-LR.pdf

20. Woodcock Language Proficiency Battery-Revised (WLPB-R)

Woodcock Language Proficiency Battery-Revised [Website]. (2015). Retrieved from https://shop.acer.edu.au/group/CG/19;jsessionid=DA68131C6325DDE9DC211F0C972EF006

Woodcock Language Proficiency Battery-Revised, (WLPB-R) English and Spanish Forms [Website]. (2015). Retrieved from http://outreach.ewu.edu/media/courses/flash/password/CEDP589/unit/html/wlpbr.htm

Schrank, F. A., Fletcher, T. V., & Alvarado, C. G. (1996). Comparative validity of three English oral language proficiency tests [PDF]. *Bilingual Research Journal, 20*, 55–68. Retrieved from http://www.ncela.us/files/rcd/be021065/comparative_validity_of_three.pdf