36

Computer-Assisted Language Testing

Ruslan Suvorov
Iowa State University

Volker Hegelheimer
Iowa State University

Introduction

Computer-assisted language testing (CALT) employs computer applications for eliciting and evaluating test takers' performance in a second language. CALT encompasses computer-adaptive testing (CAT), the use of multimedia in language test tasks, and automatic response analysis (Chapelle & Douglas, 2006). Chapelle (2010) distinguishes three main motives for using technology in language testing: efficiency, equivalence, and innovation. Efficiency is achieved through computer-adaptive testing and analysis-based assessment that utilizes automated writing evaluation (AWE) or automated speech evaluation (ASE) systems. Equivalence refers to research on making computerized tests equivalent to paper and pencil tests that are considered to be "the gold standard" in language testing. Innovation—where technology can create a true transformation of language testing—is revealed in the reconceptualization of the L2 ability construct in CALT as "the ability to select and deploy appropriate language through the technologies that are appropriate for a situation" (Chapelle & Douglas, 2006, p. 107). In addition, innovation is exemplified in the adaptive approach to test design and automatic intelligent feedback provided with the help of AWE and ASE technologies integrated in computerized tests.

Computer-based testing, once viewed as a convenient delivery vehicle for traditional paper and pencil tests (Garrett, 1991), has undergone important changes since the late 1980s. While a significant aspect of CALT continues to revolve around the delivery of paper-based tests, this area has witnessed major developments since the 1990s including computer-adaptive testing, new item types, integrated skills assessment, and automated evaluation.

In most of the recent books that deal with the assessment of various aspects of the English language (e.g., the *Cambridge Language Assessment Series*, edited by

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

DOI: 10.1002/9781118411360.wbcla083

Alderson and Bachman, on assessing vocabulary, reading, language for specific purposes [LSP], writing, listening, and grammar), the argument in favor of using computer technology to deliver assessments includes "efficacy" as a major component. That is, what is viewed as one potential advantage to using technology in assessment revolves around more expeditious test delivery, test evaluation, and score reporting. Incidentally, only *Assessing Speaking* (Luoma, 2004) does not include a section on computer technology. However, interest in the assessment of speaking has risen sharply since the early 2000s, an interest fueled in part by call center work, where speaking is viewed as a key to higher customer satisfaction ratings.

We begin this chapter by describing a framework for computer-assisted language testing that draws on scholars who have explored various aspects of CALT. We then provide a review of major computer-based tests and test delivery platforms, followed by a brief synopsis of recent research in the field. In our closing section, we outline challenges and new possibilities in CALT.

Description of Computer-Assisted Language Testing

Computer-assisted language testing comprises different aspects of language testing and technology use. In this section, we present a framework for the description of computer-assisted language tests as instruments developed within CALT (see Table 36.1) on the basis of various attributes that have been previously described in the literature, we define a computer-assisted language test as any test delivered via a computer or a mobile device. The framework consists of nine attributes and their corresponding categories. While the first five categories and the interactive category for the last attribute are unique to CALT, the remaining four attributes are also germane to traditional paper-based tests.

Table 36.1 Framework for the description of computer-assisted language tests

#	*Attribute*	*Categories*
1	Directionality	Linear, adaptive, and semi-adaptive testing
2	Delivery format	Computer-based and Web-based testing
3	Media density	Single medium and multimedia
4	Target skill	Single language skill and integrated skills
5	Scoring mechanism	Human-based, exact answer matching, and analysis-based scoring
6	Stakes	Low stakes, medium stakes, and high stakes
7	Purpose	Curriculum-related (achievement, admission, diagnosis, placement, progress) and non-curriculum-related (proficiency and screening)
8	Response type	Selected response and constructed response
9	Task type	Selective (e.g., multiple choice), productive (e.g., short answer, cloze task, written and oral narratives), and interactive (e.g., matching, drag and drop)

Directionality

Computer-assisted language testing can be linear, adaptive, or semi-adaptive. Linear tests administer the same number of test items in the same order to all test takers. In some linear tests, test takers can go back to previous questions and review their responses, whereas in other linear tests they are not allowed to do that. In computer-adaptive testing, each task is selected by the computer based on the test taker's performance on the previous task. Successful task completion results in a more complex question, while incorrect task completion results in an easier next task. By adapting the complexity of tasks to the test taker's performance, a computer-adaptive test requires ostensibly fewer items and less time to assess the language proficiency level of its users.

Unlike linear tests that often use classical test theory and its extensions, computer-adaptive tests (CATs) are based on item response theory (IRT). This test theory is based on two major assumptions: (a) unidimensionality (i.e., all test items must measure the same construct) and (b) local independence (i.e., test takers' responses to each test item must be independent from each other) (Henning, Hudson, & Turner, 1985). Depending on the type of IRT model, items for CATs can be created using one, two, or three parameters, namely, item difficulty, item discrimination, and item guessing (Jamieson, 2005).

Due to limitations of computer-adaptive testing—including high cost, increased exposure of test items, issues with algorithms for item selection, and difficulties with satisfying strict IRT assumptions—semi-adaptive tests have been proposed and used as an alternative. Compared to adaptive tests that are adaptive at the item level (i.e., by selecting the next item based on the test taker's performance on the current item), semi-adaptive tests are adaptive at the level of a group of items called testlets (Winke & Fei, 2008) or at the level of the whole test where test takers are given a version of the test that corresponds to their proficiency level as determined by a pretest (Ockey, 2009). It should be noted, however, that the term "semi-adaptive" is not universally accepted and, while some researchers distinguish semi-adaptive tests from purely adaptive tests (e.g., Winke, 2006; Ockey, 2009; Winke & Fei, 2008), others seem to consider such tests to be a specific type of adaptive test (e.g., Alderson, 2005; Jamieson, 2005).

Delivery Format

Language tests administered with the help of computers can be divided into computer-based tests (CBTs) and Web-based tests (WBTs). Computer-based testing involves the use of various offline delivery formats such as CDs, DVDs, and standalone software applications that can be installed on an individual computer. Web-based tests, on the other hand, refer to the evaluation of test takers' performance in an online format. Roever (2001) differentiates between low-tech and high-tech WBTs depending on their technological sophistication. (See Carr, 2006, for a more detailed discussion of Web-based language testing.) Some researchers (e.g., Ockey, 2009) predict that due to rapid technological advances WBT will gain more popularity and witness further development in the near future.

Media Density

One of the advantages of computer-assisted language testing regularly mentioned in the literature is the availability of different media formats and the possibility of their integration. On the basis of this attribute, tests delivered via computers can use a single medium (e.g., an audio-only listening test or a text-based reading test) or multimedia (e.g., a listening test with a video or a reading test with text and images). The use of multimedia, which may incorporate audio, images, videos, animation, and graphics, has gained much attention among researchers because it is believed to have the potential for enhancing the authenticity of language tasks. However, Douglas and Hegelheimer (2007) warn that this issue is not as straightforward as it might first seem because the implementation of multimedia in computer-assisted language tests results in a more complex construct to measure, which, in turn, poses a threat to test validity.

Target Skill

Computerized language tests can be designed to assess a single language skill (i.e., reading, writing, speaking, or listening) or a set of integrated skills (for instance, speaking and listening). Integrated skills assessment reflects the complexity of language use contexts (Chapelle, Grabe, & Berns, 2000) and is believed to enhance the authenticity of language tests through interactivity provided by integrated tasks (Ockey, 2009) that are typically performance-based (Plakans, 2009a). Integrated skills assessment, for instance, has been included in the new TOEFL iBT in order "to better align the test design to the variety of language use tasks that examinees are expected to encounter in everyday academic life" (Sawaki, Stricker, & Oranje, 2009). According to Plakans (2009a), tasks for assessing integrated skills are difficult to develop and are more prevalent in the English for specific purposes (ESP) and English for academic purposes (EAP) tests.

Scoring Mechanism

With regards to the scoring mechanism, test takers' performance on computer-delivered language tests can be evaluated by human raters and by computers. Computerized scoring of the input can be done by matching exact answers or analyzing test takers' responses. Exact answer matching entails matching test takers' responses with the correct preset responses (for instance, responses to multiple choice and matching questions). This type of scoring is typically used for the evaluation of receptive skills (i.e., reading and listening) and, sometimes, productive skills (e.g., writing) in the form of one word or even short phrase answers provided that the test has a prepiloted list of acceptable answers, including the ones with common spelling errors (Alderson, 2005). The use of analysis-based scoring, on the other hand, enables performance-based testing, where test takers construct extended responses to complete writing and speaking tasks. Analysis-based scoring utilizes various natural language processing methods integrated in many automated writing evaluation systems such as e-rater® used in Criterion (e.g., Attali & Burstein, 2006; Burstein & Chodorow, 2010) and speech

evaluation systems such as Ordinate in the Versant English Test (e.g., Downey, Farhady, Present-Thomas, Suzuki, & Van Moere, 2008). The results of such automated assessment can be provided as a holistic score, diagnostic feedback, or both (Burstein & Chodorow, 2010).

Stakes

As any type of testing, computer-assisted language testing can have low, medium, and high stakes for test takers. Low stakes testing has little, if any, consequences for test takers, and is employed for practicing, self-studying, and track-keeping purposes. Computerized tests with medium stakes (such as testing of students' progress in a second language classroom) can have some impact on test takers' lives. High stakes tests, which do have life-changing consequences and implications, are typically used for admissions to educational programs, professional certification and promotion, and granting citizenship (Roever, 2001).

Purpose

Test purpose is associated with the type of tests and decisions that can be made on the basis of the test performance. Carr (2011) classifies test purposes into two broad categories: curriculum-related and other, or non-curriculum-related (p. 6). Curriculum-related tests can be used for the purposes of admission to a program, placement into a specific level of the program, diagnosis of test takers' strengths and weaknesses, assessment of their progress in the program, and their achievement of the program's objectives. The non-curriculum-related tests are used for language proficiency assessment and screening for non-academic purposes (e.g., to make decisions regarding employment, immigration, etc.).

Response Type

There are two main types of responses that can be provided by test takers during a computer-delivered language test: selected and constructed responses (e.g., Parshall, Davey, & Pashley, 2000). Selected response assessment involves tasks that require a test taker to choose a correct answer from a list of options (e.g., a multiple choice question). In the case of constructed responses, test takers must develop their own answers and produce short or extended linguistic output. These two categories, however, should be viewed continuously rather than dichotomously since some language tasks can require a response that would possess the features of both (for instance, arranging given words and phrases into a sentence).

Task Type

There are numerous types of tasks that can be created for computerized language tests. Task types can be divided into three broad categories: selective (e.g., multiple choice questions, yes/no questions), productive (e.g., written and oral narratives, short answer tasks, and cloze tasks), and interactive (e.g., matching, drag and

drop). Although some of these tasks are also possible in a paper and pencil test, others can be created and delivered only through computers. Alderson (2005), for instance, describes 18 experimental items that were created as part of the DIALANG project, which is a low stakes computer-based diagnostic test available in 14 European languages. According to Alderson (2005), these innovative items provide new opportunities for enhanced diagnosis and alternative types of feedback in CALT. Some examples of these items include multimedia-enriched items (e.g., pictorial multiple choice with sound, interactive image with sound, and video clips in listening), interactive items that require test takers to manipulate the test content (e.g., reorganization, highlighting/underlining, insertion, deletion, thematic grouping, transformation), and items that provide alternative ways to assess productive skills (e.g., indirect speaking with audio clips as alternatives, benchmarking in direct writing, and multiple benchmarks in speaking).

There is an obvious interaction among all the attributes from Table 36.1. Test purpose, for instance, may be interrelated with target skills: Diagnostic tests that are "more likely to be discrete-point than integrative" (Alderson, 2005, p. 11) tend to focus on a specific language skill (e.g., reading), whereas in proficiency tests the assessment of integrated skills is more preferable. Likewise, stakes may affect the selection of a scoring mechanism and delivery format: Considering the existing limitations of automated evaluation systems and potential risks associated with Web-based testing, high stakes test developers will likely opt for the CBT format and combine analysis-based scoring with human-based scoring, while some low stakes tests may welcome WBT and rely exclusively on automated assessment.

Computer-Based Tests and Delivery Platforms

Rapid technological advances and the ensuing quick expansion of computer-assisted language testing have resulted in a variety of commercial computer-delivered language tests and platforms for creating customized assessment. Hence, the discussion in this section will be divided into two streams: existing computerized language tests and instruments for constructing original L2 tests.

Existing Computerized L2 Tests

Since the emergence of the first computer-based and computer-adaptive language tests in the 1980s, numerous CALT projects have been initiated by academic institutions and test development companies. Chalhoub-Deville (2010) reviews a representative sample of computer-delivered language tests discussed in the research literature over the past several decades. This section, however, will briefly describe only the most recent and innovative developments in CALT that have gone beyond a simple adaptation of paper and pencil tests for computer delivery. Specifically, we will provide a short overview of major language tests that include the assessment of productive skills, mention their purpose and structure, and focus on some of their technology-enabled features such as automated scoring algorithms, the adaptive approach, and innovative test items.

Test of English as a Foreign Language Internet-Based Test (TOEFL iBT®) Being a high stakes test, TOEFL® (published by the Educational Testing Service, http://www.ets.org/toefl/ibt/about/) is probably one of the most recognized and known language tests in the world. First introduced in 2005, TOEFL iBT witnessed several major changes compared to the older, computer-based version of TOEFL (i.e., TOEFL CBT). In particular, the adaptive approach that was used in the structure and listening sections of TOEFL CBT was discontinued in the new TOEFL, whereas a new type of tasks—called integrated tasks—was introduced. Since the use of integrated tasks violated the assumptions of a three-parameter IRT model used in the adaptive part of TOEFL CBT, ETS made a decision to abandon the adaptive approach. This decision was also prompted by the need to have human raters assess test takers' speaking and writing responses (Jamieson, Eignor, Grabe, & Kunnan, 2008).

The purpose of TOEFL iBT is to measure the ability of non-native speakers of English to perform university-level academic tasks using their English language skills. Although TOEFL iBT scores are used primarily by English-medium universities around the world for making admission decisions, they are also accepted by immigration departments and various licensing agencies. The whole test lasts for about 4.5 hours and consists of four main sections: reading, listening, speaking, and writing. Integrated tasks include reading a text, listening to a lecture or a conversation, and providing a written or an oral response on the basis of what has been read and heard. The writing section of TOEFL iBT is evaluated by human raters and an automated scoring system called e-rater. The speaking section of a practice exam for the TOEFL is evaluated by the SpeechRater℠ engine; however, this automated scoring system is not used in the actual test (Higgins, Zechner, Xi, & Williamson, 2011).

BULATS Online Tests Business Language Testing Service (BULATS) online tests (published by Cambridge ESOL, http://www.bulats.org/Bulats/The-Tests.html) comprise the BULATS Online Reading and Listening Test, BULATS Online Speaking Test, and BULATS Online Writing Test. Designed to test the English language proficiency of business employees, job applicants, and candidates for business English language courses, these three high stakes tests can be used separately or in any combination depending on the client's assessment needs.

The BULATS Online Reading and Listening Test utilizes the adaptive approach to item selection, presenting new tasks on the basis of test takers' responses to the previous items. The test consists mainly of multiple choice questions and, depending on the level of test takers' language proficiency, lasts for about an hour. Individual scores for reading and listening as well as an overall score are calculated and displayed immediately after the completion of this test (Cope, 2009).

BULATS Online Speaking includes practical tasks that require test takers to answer interview questions, read aloud sentences, give two 1-minute presentations, and express their opinions on a topic. Responses are recorded on a computer and later evaluated by human raters.

Finally, the 45-minute BULATS Online Writing test assesses Business English writing skills via two tasks that must be completed on a computer in response to

given prompts: a 50–60-word e-mail and a 180–200-word report. Responses are subsequently rated by trained examiners.

BEST Plus™ Computer-Adaptive Version Basic English Skills Test (BEST) Plus (published by the Center for Applied Linguistics, http://www.cal.org/aea/best-plus/ca.html) is designed to assess the listening and speaking skills of adult learners of English in the US context. The computer-adaptive version of this test is CD-ROM based and takes 3 to 20 minutes to complete, depending on test takers' oral skills. There are seven types of tasks on various general topics such as health, transportation, and housing. The item types comprise photo description, entry item, yes/no question, choice question, personal expansion, general expansion, and elaboration. Upon reading a task to the candidate from the computer screen, a trained test administrator instantly evaluates the candidate's response and enters the score in the computer. The answers are scored on the basis of listening comprehension, language complexity, and communication (Van Moere, 2009). The next item selected by the BEST Plus system is based on the test taker's response to the previous question. Test scores are generated by the computer and become available immediately after the test.

COMPASS® ESL Placement Test The main purpose of the COMPASS ESL Placement Test (published by ACT, http://www.act.org/compass/tests/esl.html) is to assess the standard American English language skills of ESL students and place them into appropriate ESL courses at post-secondary educational institutions in the USA. The four major components of this computer-adaptive test— ESL Listening, ESL Reading, ESL Grammar/Usage, and ESL Essay (ESL e-Write)—can be administered either separately or in any combination.

The first three parts of the COMPASS ESL Placement Test are composed mostly of multiple choice questions (with some modified cloze items in the ESL Grammar/Usage test) that derive from listening and reading passages on various academic topics. The adaptive format of the test adjusts the difficulty level of the selected items to the individual test taker's performance. Based on the separate scores for the ESL Listening, ESL Reading, and ESL Grammar/Usage tests, students are assigned one of the four levels.

The 30-minute ESL Essay test is delivered and assessed online using automated scoring technology. The overall score for this test is assigned on a six-point scale and incorporates analytic scores for development, focus, organization, language use, and mechanics.

Versant™ English Test The Versant English Test (published by Pearson, http://www.versanttest.com/products/english.jsp), formerly known as PhonePass and Spoken English Tests (SET-10), is an automated test designed to measure the English speaking skills of non-native English speakers. This high stakes test is used in education and business for admission, recruitment, and promotion purposes.

The Versant English Test is composed of six sections: reading, repeats, short answer questions, sentence builds, story retelling, and open questions. It lasts for approximately 15 minutes and can be delivered over a telephone or a computer,

with tasks being presented orally in native-sounding voices. The assessment of test takers' responses is done by an automated speech evaluation system called Ordinate that assigns scores within several minutes after test completion. In addition to an overall score, test takers also receive individual subscores for sentence mastery, vocabulary, fluency, and pronunciation.

Pearson Test of English (PTE) Academic™ Developed by the same publisher as the Versant English Test, PTE Academic (Pearson, http://www.pearsonpte.com/pteacademic) is designed to measure the English language proficiency of international students in the academic context. First introduced in 2009, this high stakes computer-based test lasts for three hours and consists of four parts: introduction, speaking and writing, reading, and listening. According to the publisher, PTE Academic uses 20 innovative item types including items that provide integrated skills assessment. The test employs automated scoring tools to assess test takers' productive skills: The Intelligent Essay Assessor™ (IEA) is used to evaluate writing skills, whereas Pearson's Ordinate technology is integrated in the assessment of speaking. Score reports, consisting of an overall score, scores for communicative skills (i.e., listening, speaking, reading, and writing), and scores for enabling skills (i.e., grammar, spelling, pronunciation, oral fluency, vocabulary, and written discourse), are available online within five days of test completion. Each score ranges from 10 to 90 points. To date, PTE Academic appears to be the only high stakes computerized language test that uses automated assessment of both productive skills.

L2 Test Development Instruments

The advent of emerging technologies and the Web 2.0 era has generated a number of tools that can be utilized by language educators and practitioners for the development and delivery of low and medium stakes L2 assessment. These instruments include both standalone virtual learning environments that, among other educational purposes, can be used for creating and administering computer-based language tests, and specialized applications to construct individual test items that can later be embedded in different delivery platforms. In this section, we will adumbrate the principal features of the major free (Moodle and Google Docs) and commercial (Respondus and Question*mark* Perception) options for creating computer-based language tests.

Moodle 2.2 Moodle (http://moodle.org/) is designed for teaching and learning purposes in a variety of educational settings. The latest version of this open-source course management system (CMS), released on December 5, 2011, provides some advanced opportunities for testing and assessment. The new Moodle 2.2 question bank allows for the creation of both selected response items (e.g., true/false, multiple choice, and matching questions) and constructed response items (e.g., cloze, short answer, and essay questions). Latest features in the question bank include new feedback options and delivery modes for presenting questions to test takers: adaptive mode, interactive mode, deferred feedback, immediate feedback, and manual grading. Besides the built-in quiz module that enables the integration of

questions from the question bank and provides various reports statistics, language instructors can utilize third party quiz modules for Moodle such as TaskChain (formerly QuizPort) that come with more advanced assessment features. TaskChain, for example, can be used to create semiadaptive tests that consist of an optional entry Web page followed by a set of quizzes with multimedia content and an optional exit Web page (see http://docs.moodle.org/20/en/QuizPort_ module for more information about TaskChain).

Google Docs This Web-based office suite (https://docs.google.com) is a good free solution for Google users who want to easily create and publish online quizzes and tests using a Google form. This free application supports several types of questions including multiple choice, checkboxes, text (short answer questions), paragraph text (extended answer questions), and choose from a list questions. To make assessment more visually appealing to test takers, Google provides dozens of customizable themes that can be applied to tests and quizzes. Tests can be delivered via emails or embedded in other Web pages. Once students have completed the assessment, Google Docs will immediately generate reports with students' responses and summarize the results in a graphic form. More advanced features include the implementation of formulas to automatically calculate the number of correct points and final grades received by test takers.

Respondus® This commercial assessment tool (http://www.respondus.com) is designed for the development of tests that can be integrated in various learning management systems such as Moodle, Blackboard, ANGEL, and Desire2Learn. Respondus supports 15 question types including multiple choice, true/false, matching, short answer, and paragraph-writing tasks. Moreover, this application allows for the use of images, sound, video, and embedded Web content, thus offering language instructors a great degree of flexibility and helping enhance the authenticity of tests. The results of delivered assessments can be saved as custom reports and downloaded in an Excel format. Other options in Respondus include easy archiving and restoration as well as key word search to locate specific questions within a test.

*Question*mark™ *Perception*™ Question*mark* Perception (http://www.questionmark.com/us/perception) is an assessment management system conceived as a tool for educators and evaluation experts to create and deliver different types of tests, quizzes, and exams. Similar to Respondus, this system supports publishing of tests in other learning management systems using SCORM packages. Question*mark* Perception can be used to create a great variety of question types. Some innovative items that might be of interest to language testing professionals include Captivate Simulation that utilizes simulation questions created in Adobe Captivate, and Spoken Response that allows test takers to record their responses in an audio format. In addition to multimedia and Flash support, Question*mark* Perception provides options for importing questions from ASCII, QML, and QTI XML files. Assessments created with the help of this system can be delivered through

standard Web interface and applications for mobile devices. To prevent cheating, questions in high stakes tests can be administered in a secure mode through a Question*mark* secure server.

This review of existing commercial products as well as tools used for the development of customized computer-based language tests reveals a variety of available language assessment options. While these assessment options demonstrate a strong potential of technology, they also expose challenges in computer-assisted language testing, including the difficulty of providing automated feedback on speaking tests and conducting fully automated evaluation of essays. Many of these challenges are the focal point of present research in CALT. Current efforts also revolve around conducting construct validation research, creating new types of tasks, integrating multimedia in increasingly more authentic language tasks, and advancing integrated skills assessment.

Research Studies and Major Developments in CALT

Construct Validity and Comparability Studies

A great deal of research in the field of CALT has been dedicated to investigating construct validity of computer-based tests. Construct validity evidence refers to "the judgmental and empirical justifications supporting the inferences made from test scores" (Chapelle, 1998, p. 50). According to Dooey (2008), construct validation in CALT is of utmost importance because it helps ensure that the test is measuring test takers' specific language skill(s) rather than their computer skills. Such validation can be done by comparing traditional (i.e., paper-based) and computer-based language tests. Although comparability studies are often commissioned by test development companies, more independent research comparing paper-based and computer-based language tests is also available (e.g., Sawaki, 2001; Coniam, 2006).

One such independent comparability study was conducted by Sawaki (2001), who reviewed the assessment literature to examine the equivalence between conventional and computerized L2 reading tests. This yielded mixed empirical findings vis-à-vis the comparability of paper-based and computer-based L2 reading tests, highlighting the dearth of research on the effect of the mode of presentation on L2 reading and limitations of the methodological approaches used in the existing studies. The results of Coniam's (2006) study, however, appeared to be more conclusive. He found high correlation between test takers' performance on computer-based and paper-based L2 listening tests, even though their scores on the CBT appeared higher than on the conventional test.

Development of Computerized Language Tests

Another line of research in CALT focuses on reporting the development of CATs or other CBTs. Despite the large number of existing commercial assessments, individual researchers and institutions pursue the development of "homemade"

language tests to match their specific needs. Papadima-Sophocleous (2008), for instance, reports on the development of a computer-based online test, NEPTON, that attempts to combine the advantages of CBTs and CATs. The items for this test were selected on the basis of both content and statistical properties (e.g., item difficulty), and target different language competence levels, language skills, and activity types. Unlike a typical computer-adaptive language test, NEPTON allows test takers to browse the questions, change the responses, and complete the questions in any order.

Two other customized computer-based language tests are discussed in the studies by Alderson and Huhta (2005) and Roever (2006). Alderson and Huhta (2005) describe the development of a Web-based language assessment system called DIALANG. This large-scale project involved 25 higher education institutions in the European Union. Based on the Common European Framework of Reference (CEFR), DIALANG provides diagnostic assessment of reading, writing, listening, vocabulary, and grammar in 14 different European languages. Due to the computer-adaptive nature of the test, test takers are given the version of the test based on their responses to the vocabulary test and self-evaluation statements that they have completed at the beginning of the assessment. Another feature of DIALANG is its detailed feedback coupled with suggestions for test takers on how to move to the next CEFR level (Alderson & Huhta, 2005).

The test reported in Roever's (2006) study is a Web-based test of ESL pragmalinguistics. Consisting of 36 multiple choice and short answer items, this low stakes assessment was designed to measure the ESL learners' knowledge of speech acts, implicature, and routines. Although, according to Roever (2006), the test was sufficiently reliable and proved that it was possible to evaluate L2 knowledge of pragmalinguistics, it did not assess users' knowledge of sociopragmatics and relied on the written format for the evaluation of the speech act responses.

Use of Multimedia in CALT

The use of multimedia in computer-delivered language tests has been the focus of debate since the early 1990s. Some experts suggest that the inclusion of multimedia in language tests "can assist us in simulating a great many aspects of communicative language use situations" (Douglas, 2010, p. 118), thus making such tests more authentic. However, research in this area, namely on the use of visuals for listening assessment, has yielded some contentious results. On one hand, the use of multimedia has been found facilitative for test takers' performance on L2 listening tests (e.g., Ginther, 2002). Findings of other studies, however, suggest that test takers can get distracted by video and images (Wagner, 2007; Suvorov, 2009). According to Fulcher (2003), the integration of multimedia in speaking tests is even more problematic due to challenges with the timing of test takers' recordings on one hand, and the dearth of research on the effect of visuals on test takers' performance on the other hand. Thus, the question of whether it is worth investing the time and money to create and implement multimedia in language tests remains open.

Integrated Skills Assessment

Another trend of CALT research focuses on integrated skills assessment. Unlike the testing of unitary skills such as speaking, listening, reading, and writing, this type of assessment is believed to be more authentic due to the interactive nature of tasks that resemble what test takers may encounter in real-world situations (Jamieson, 2005; Ockey, 2009). Several major language tests have recently implemented tasks that assess the integrated skills of speaking and listening (Versant English Test); reading–listening–writing and listening/reading–speaking (TOEFL iBT); and reading–writing, listening–writing, listening–speaking, and reading–speaking (PTE Academic). Although Ockey (2009) maintains that "the future of integrated skills tests appears bright" (p. 845), the use of integrated tasks in CBTs poses certain challenges, namely the vagueness of language ability constructs being measured by such tests. This demands more research on multidimensional constructs and on inferences that can be made about test takers' language proficiency based on their scores for integrated items (Plakans, 2009b).

Research on Automated Assessment

Significant research efforts are being employed in the area of automated assessment of productive skills. Although automated evaluation has been in use for an extended period of time, its application in language assessment is relatively new (Chapelle & Chung, 2010). Research on automated writing evaluation has resulted in products such as Intelligent Essay Assessor (Pearson), e-rater (ETS), and IntelliMetric® (Vantage Learning) that are capable of analyzing lexical measures, syntax, and discourse structure of essays. The Intelligent Academic Discourse Evaluator (IADE) is another example of a Web-based AWE program that utilizes NLP techniques to provide feedback at the level of rhetorical functions in research writing (Cotos, 2011). IADE has become the prototype of a complex AWE system currently under development at Iowa State University.

Although AWE systems are used extensively in many educational institutions, these systems are not universally accepted. According to Cotos (2011), supporters suggest that AWE systems are generally in close agreement with human raters and are thus more time- and cost-effective. They may also foster learner autonomy, promote the process writing approach that involves writing multiple drafts, and lead to individualized assessment. Critics, however, claim that the use of such systems encourages students to focus on surface features such as grammar and vocabulary rather than meaning. In addition, automated assessment of essays diminishes the role of instructors and impels students to adjust their writing to the evaluation criteria of these systems (Cotos, 2011).

Unlike automated writing assessment, ASE involves an additional layer of complexity in that the test takers' oral output must first be recognized before it can be evaluated (Xi, 2010a). Despite ongoing research and recent advancements in automated speech recognition (ASR), these technologies are not robust at recognizing non-native accented speech because most ASR-based systems have been designed for a narrow range of native speech patterns. This limitation has been addressed in CALT in two ways. First, some automated speech

evaluation systems (e.g., the one used in the Versant speaking tests developed by Pearson) constrain the context of the utterance so that users' spoken output becomes highly predictable. Other ASE systems (e.g., SpeechRater developed by ETS) compensate for this limitation with free speech recognition by expanding the speaking construct to include pronunciation, vocabulary, and grammar, in addition to fluency (Xi, Higgins, Zechner, & Williamson, 2008). According to Xi (2010a), currently neither of these approaches "has successfully tackled the problem of under- or misrepresentation of the construct of speaking proficiency in either the test tasks used or the automated scoring methodologies, or both" (p. 294).

As shown in this section, research in CALT has led to several major developments, including multimedia language tasks, integrated skills assessment, and automated evaluation of productive skills. Although many of these developments have made a significant impact on language assessment, some of them showed only the potential promise of technology for advancing the field of computer-assisted language testing.

Challenges and New Possibilities in CALT

The views regarding the current status and the future of CALT vary slightly among researchers, with some being more concerned about the severity of existing problems than others. Ockey (2009), for instance, believes that due to numerous limitations and problems "CBT has failed to realize its anticipated potential" (p. 836), while Chalhoub-Deville (2010) contends that "L2 CBTs, as currently conceived, fall short in providing any radical transformation of assessment practices" (p. 522). In the meantime, other researchers (e.g., Chapelle, 2010; Douglas, 2010) appear to be somewhat more positive about the transformative role of CALT and stress that despite existing unresolved issues technology remains "an inescapable aspect of modern language testing" and its use in language assessment "really isn't an issue we can reasonably reject—technology is being used and will continue to be used" (Douglas, 2010, p. 139).

Still, everyone seems to acknowledge the existence of challenges in CALT, maintaining that more work is necessary to solve the persisting problems. In particular, a noticeable amount of discussion in the literature has been dedicated to the issues plaguing computer-adaptive testing, which, according to some researchers, led to the decline of its popularity, especially in large scale assessment (e.g., Douglas & Hegelheimer, 2007; Ockey, 2009). Of primary concern for CATs is the security of test items (Wainer & Eignor, 2000). Unlike a linear CBT that presents the same set of tasks to a group of test takers, a computer-adaptive language test provides different questions to test takers. To limit the exposure of items, CATs require a significantly larger item pool, which makes the construction of such tests more costly and time-consuming. Ockey (2009) suggests that one way to avoid problems associated with test takers' memorization of test items is to create computer programs that would generate questions automatically.

Furthermore, there is no agreement on which algorithm to use for selecting items in CATs (Ockey, 2009). Some test developers suggest starting a CAT with

easy items, whereas others recommend beginning with items of average difficulty. Additionally, no consensus has been reached on how the algorithm should proceed with the selection of items once a test taker has responded to the first question, nor are there agreed-upon rules on when exactly an adaptive test should stop (Thissen & Mislevy, 2000). Nonetheless, research is being carried out to address this issue and new methods of item selections in computer-adaptive testing such as the Weighted Penalty Model (see Shin, Chien, Way, & Swanson, 2009) have recently been proposed.

Another major problem with computer-adaptive tests concerns their reductionist approach to the measured L2 constructs. Canale (1986) was one of the first to argue that the unidimensionality assumption deriving from the IRT models used in CATs poses a threat to the L2 ability construct, making it unidimensional as well. This concern has further been reiterated by other experts in language assessment (e.g., Chalhoub-Deville, 2010; Douglas, 2010). Their main argument suggests that the L2 ability construct should be multidimensional and consist of multiple constituents that represent not only the cognitive aspects of language use, but also knowledge of language discourse and the norms of social interaction, the ability to use language in context, the ability to use metacognitive strategies, and, in the case of CALT, the ability to use technology. Hence, Chalhoub-Deville (2010) asserts that, because of the multidimensional nature of the L2 ability construct, measurement models employed in CBTs must be multidimensional as well—a requirement that many adaptive language tests do not meet. Finally, the unidimensionality assumption of IRT also precludes the use of integrated language tasks in computer-adaptive assessment (Jamieson, 2005). As a result of some of these problems, ETS, for instance, decided to abandon the computer-adaptive mode that was employed in TOEFL CBT and instead return to the linear approach in the newer TOEFL iBT.

The limitations of the adaptive approach prompted some researchers to move toward semiadaptive assessment (e.g., Winke, 2006). The advantages of this type of assessment include a smaller number of items (compared to linear tests) and the absence of necessity to satisfy IRT assumptions. Thus, Ockey (2009) argues that semiadaptive tests can be the best compromise between adaptive and linear approaches and predicts that they will become more widespread in medium-scale assessments.

Automated scoring is another contentious area of CALT. One of the main issues with automated scoring of constructed responses, both for writing and for speaking assessment, is related to the fact that computers look only at a limited range of features in test takers' output. Even though research studies report relatively high correlation indices between the scores assigned by AWE systems and human raters (e.g., Attali & Burstein, 2006), Douglas (2010) points out that it is not clear whether the underlying basis for these scores is the same. Specifically, he asks, "are humans and computers giving the same score to an essay but for different reasons, and if so, how does it affect our interpretations of the scores?" (Douglas, 2010, p. 119). He thus concludes that although "techniques of computer-assisted natural language processing become more and more sophisticated, . . . we are still some years, perhaps decades, away from being able to rely wholly on such systems in language assessment" (Douglas, 2010, p. 119). Since machines do not

understand ideas and concepts and are not able to evaluate the meaningful writing, critics contend that AWE "dehumanizes the writing situation, discounts the complexity of written communication" (Ziegler, 2006, p. 139) and "strikes a death blow to the understanding of writing and composing as a meaning-making activity" (Ericsson, 2006, p. 37).

Automatic scoring of speaking skills is even more problematic than that of writing. In particular, speaking assessment involves an extra step which writing assessment does not have: recognition of the input (i.e., speech). Unlike writing assessment, the assessment of speaking also requires the evaluation of segmental features (e.g., individual sounds and phonemes) and suprasegmental features (e.g., tone, stress, and prosody). Since automated evaluation systems cannot perform at the level of human raters and cannot evaluate coherence, content, and logic the way humans do, they are used almost exclusively in conjunction with human raters. As Xi (2010b) concludes, "We are not ready yet to use automated scoring alone for speaking and writing in high-stakes decisions."

Other challenges faced by CALT are related to task types and design, namely the use of multimedia and integrated tasks. Although the use of multimedia input is believed to result in a greater level of authenticity in test tasks by providing more realistic content and contextualization cues, it remains unclear how the inclusion of multimedia affects the L2 construct being measured by CBTs (Jamieson, 2005). Some researchers even question the extent to which multimedia enhances the authenticity of tests (e.g., Douglas & Hegelheimer, 2007) since comparative studies on the role of multimedia in language assessment have yielded mixed results (see Ginther, 2002; Wagner, 2007; Suvorov, 2009). With regards to integrated tasks, their implementation in CBTs is generally viewed favorably because such tasks seem to better reflect what test takers would be required to do in real-life situations. The use of integrated tasks is therefore believed to increase authenticity of language tests (Fulcher & Davidson, 2007). However, Douglas (2010) warns that the interpretation of integrated tasks can be problematic because, if the test taker's performance is inadequate, it is virtually impossible to find out whether such performance is caused by one of the target skills or their combination. This concern appears to be more relevant in high stakes testing than in low stakes testing.

Despite all the above-mentioned issues and concerns, most experts in computer-assisted language testing agree that technological advances and innovative measurement models will move this field forward and "the world of CALT will continue to develop" (Winke & Fei, 2008, p. 362). For true innovations and transformation of technology-enhanced language assessment to occur, CALT must be reconceptualized through "fundamental changes in the representation of the L2 construct, overall test design, task development, and even the context and purpose of tests" (Chalhoub-Deville, 2010, p. 522). New possibilities for CALT include, but are not limited to, integrating CBTs in distance and online language education; creating computer-based tests for narrower, more specific purposes; exploring the potential of technology (for instance, eye-tracking systems that enable screen navigation through eye movements) for designing language tests that will be able to better accommodate test takers with disabilities; developing innovative, more

authentic test items; and conducting interdisciplinary research to advance the field of automated scoring. Progress in automatic speech recognition and emotion recognition systems that identify emotions from speech using facial expressions, voice tone, and gestures (see Schuller, Batliner, Steidl, & Seppi, 2009) will inevitably create new opportunities for computer-based assessment of speaking. Furthermore, with the anticipated advent of Web 3.0 (Semantic Web), where computers will be able to generate new information, computer-assisted language testing might gradually evolve to the point where test items will be automatically generated by computers. For instance, to make speaking tests more authentic and mimic real-life situations, computers will act both as raters and as interlocutors, creating new tasks based on students' responses and adapting these tasks to students' performance. In the meantime, regardless of the types of future transformations and innovations that will occur in CALT, we should never forget Douglas's (2000) warning that "language testing . . . driven by technology, rather than technology being employed in the services of language testing, is likely to lead us down a road best not traveled" (p. 275).

SEE ALSO: Chapter 13: Assessing Integrated Skills; Chapter 19: Tests of English for Academic Purposes in University Admissions; Chapter 60: New Media in Language Assessments; Chapter 64: Computer-Automated Scoring of Written Responses; Chapter 75: Item Response Theory in Language Testing; Chapter 94: Ongoing Challenges in Language Assessment; Chapter 99: Assessing English in the Middle East and North Africa

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, England: Continuum International Publishing.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, *22*(3), 301–20.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, *4*(3), 1–31.

Burstein, J., & Chodorow, M. (2010). Progress and new directions in technology for automated essay evaluation. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed., pp. 529–38). Oxford, England: Oxford University Press.

Canale, M. (1986). The promise and threat of computerised adaptive assessment of reading comprehension. In C. W. Stansfield (Ed.), *Technology and Language Testing* (pp. 29–46). Washington, DC: TESOL.

Carr, N. T. (2006). Computer-based testing: Prospects for innovative assessment. In L. Ducate & N. Arnold (Eds.), *Calling on CALL: From theory and research to new directions in foreign language teaching (CALICO monograph series, 5*, pp. 289–312). San Marcos, TX: CALICO.

Carr, N. (2011). *Designing and analyzing language tests*. Oxford, England: Oxford University Press.

Chalhoub-Deville, M. (2010). Technology in standardized language assessments. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed., pp. 511–26). Oxford, England: Oxford University Press.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Second language acquisition and language testing interfaces* (pp. 32–70). Cambridge, England: Cambridge University Press.

Chapelle, C. A. (2010). *Technology in language testing* [video]. Retrieved November 14, 2012 from http://languagetesting.info/video/main.html

Chapelle, C. A., & Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, *27*(3), 301–15.

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

Chapelle, C., Grabe, W., & Berns, M. (2000). *Communicative language proficiency: Definition and implications for TOEFL 2000. TOEFL monograph series 10*. Princeton, NJ: Educational Testing Service.

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCALL*, *18*(2), 193–211.

Cope, L. (2009). CB BULATS: Examining the reliability of a computer-based test. *Research Notes*, *38*, 31–4.

Cotos, E. (2011). Potential of automated writing evaluation feedback. *CALICO Journal*, *28*(2), 420–59.

Dooey, P. (2008). Language testing and technology: Problems of transition to a new era. *ReCALL*, *20*(1), 21–34.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.

Douglas, D. (2010). *Understanding language testing*. London, England: Hodder Education.

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics*, *27*, 115–32.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly*, *5*, 160–7.

Ericsson, P. (2006). The meaning of meaning: Is a paragraph more than an equation? In P. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 28–38). Logan: Utah State University Press.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*, *20*(4), 384–408.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London, England: Routledge.

Garrett, N. (1991). Technology in the service of language learning: Trends and issues. *Modern Language Journal*, *75*, 74–101.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, *19*(2), 133–67.

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing*, *2*(2), 141–54.

Higgins, D., Zechner, K., Xi, X., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, *25*(2), 282–306.

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, *25*, 228–42.

Jamieson, J., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). The frameworks for the reconceptualization of TOEFL. In C. Chapelle, J. Jamieson & M. Enright (Eds.), *The new TOEFL* (pp. 55–95). Mahwah, NJ: LEA.

Luoma, S. (2004). *Assessing speaking. Cambridge language assessment series*. Cambridge, England: Cambridge University Press.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal*, *93*, 836–47.

Papadima-Sophocleous, S. (2008). A hybrid of a CBT- and a CAT-based New English Placement Test Online (NEPTON). *CALICO Journal*, *25*(2), 276–304.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. J. Van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129–48). Dordrecht, Netherlands: Kluwer.

Plakans, L. (2009a). *Integrated assessment* [video]. Retrieved November 14, 2012 from http://languagetesting.info/video/main.html

Plakans, L. (2009b). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561–87.

Roever, C. (2001). Web-based language testing. *Language Learning & Technology*, 5(2), 84–94.

Roever, C. (2006). Validation of a Web-based test of ESL pragmalinguistics. *Language Testing*, *23*(2), 229–56.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology*, *5*(2), 38–59.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*(1), 5–30.

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2009). Emotion recognition from speech: Putting ASR in the loop. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, 4585–8.

Shin, C. D., Chien, Y., Way, W. D., & Swanson, L. (2009). *Weighted penalty model for content balancing in CATs*. Retrieved November 14, 2012 from http://education.pearsonassessments.com/NR/rdonlyres/99A4327B-5968-4AB2-A8CD-8D502D22C2DE/0/WeightedPenaltyModel.pdf

Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. A. Chapelle, H. G. Jun, & I. Katz (Eds.), *Developing and evaluating language learning materials* (pp. 53–68). Ames: Iowa State University.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101–33). Mahwah, NJ: Erlbaum.

Van Moere, A. (2009). Test review: BEST Plus Spoken Language Test. *Language Testing*, *26*(2), 305–13.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology*, *11*(1), 67–86.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 271–99). Mahwah, NJ: Erlbaum.

Winke, P. (2006). Online assessment of foreign language proficiency: Meeting development, design, and delivery challenges. In S. Howell & M. Hricko (Eds.), *Online assessment and measurement: Case studies from teacher education, K-12 and corporate* (pp. 82–97). London, England: Information Science Publishing.

Winke, P., & Fei, F. (2008). Computer-assisted language assessment. In N. Van Deusen-Scholl & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (Vol. 4, pp. 353–64). New York, NY: Springer.

Xi, X. (2010a). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291–300.

Xi, X. (2010b). *Automated scoring* [video]. Retrieved November 14, 2012 from http://languagetesting.info/video/main.html

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0 (ETS research report no. RR-08-62)*. Princeton, NJ: Educational Testing Service.

Ziegler, W. (2006). Computerized writing assessment: Community college faculty find reasons to say "not yet." In P. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 138–46). Logan: Utah State University Press.

Suggested Readings

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology*, *1*(1), 44–59.

Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language Learning & Technology*, *5*(2), 95–8.

Chalhoub-Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, *19*, 273–99.

Noijons, J. (1994). Testing computer-assisted language testing: Towards a checklist for CALT. *CALICO Journal*, *12*(1), 37–58.