Utilizing Technology in Language Assessment

Carol A. Chapelle and Erik Voss

Abstract

This entry presents an overview of the past, present, and future of technology use in language assessment, also called computer-assisted language testing (CALT), with a focus on technology for delivering tests and processing test takers' linguistic responses. The past developments include technical accomplishments that contributed to the development of computer-adaptive testing for efficiency, visions of innovation in language testing, and exploration of automated scoring of test takers' writing. Major accomplishments include computer-adaptive testing as well as some more transformational influences for language testing: theoretical developments prompted by the need to reconsider the constructs assessed using technology, natural language-processing technologies used for evaluating learners' spoken and written language, and the use of methods and findings from corpus linguistics. Current research investigates the comparability between computer-assisted language tests and those delivered through other means, expands the uses and usefulness of language tests through innovation, seeks high-tech solutions to security issues, and develops more powerful software for authoring language assessments. Authoring language tests with ever changing hardware and software is a central issue in this area. Other challenges include understanding the many potential technological influences on test performance and evaluating the innovations in language assessment that are made possible through the use of technology. The potentials and challenges of technology use in language testing create the need for future language testers with a strong background in technology, language testing, and other areas of applied linguistics.

C.A. Chapelle (✉)
Applied Linguistics, Department of English, Iowa State University, Ames, IA, USA
e-mail: carolc@iastate.edu

E. Voss
NU Global, Northeastern University, Boston, MA, USA
e-mail: e.voss@neu.edu

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_10

Keywords
Assessment • Computer • Technology • Testing

Contents

Introduction

Technology is often associated with efficiency. Accordingly, applied linguists might consider technology in language assessment in terms of how it streamlines the testing process. Indeed, much progress can be identified with respect to this worthwhile goal, as many language tests today are delivered by computer to increase efficiency. An equally important strand of language assessment concerns the relationship of language assessment to language learning, language teaching, and knowledge within the field of applied linguistics. The story of technology in language assessment needs to encompass both the efficiency of technical accomplishments and the ways that these tests intersect with other factors in the educational process for language learners. Technology can include a broad range of devices used in the testing process, from recording equipment, statistical programs, and databases to programs capable of language recognition (Burstein et al. 1996). However, here the focus will be on the use of computer technology for delivering tests and processing test takers' linguistic responses because these are the practices with the most direct impact on test takers and educational programs. The use of computer technology in language assessment is referred to as computer-assisted language assessment or computer-assisted language testing (CALT), two phrases that are used interchangeably.

Early Developments

Early developments in computer-assisted language assessment consisted of a few demonstration projects and tests used in university language courses. Many of these were reported in two edited collections, *Technology and Language Testing* (Stansfield 1986) and *Computer-Assisted Language Learning and Testing: Research Issues and Practice* (Dunkel 1991), but others had been published as journal articles. Three important themes were prevalent in this early work.

One was the use of a psychometric approach called item response theory (Hambleton et al. 1991), which provides a means for obtaining robust statistical data on test items. These item statistics, obtained from pretesting items on a large group of examinees, are used as data by a computer program to help select appropriate test questions for examinees during test taking. Item response theory, which offers an alternative to calculation of item difficulty and discrimination through classical true score methods, entails certain assumptions about the data. The use of these methods, the assumptions they entail, and the construction and use of the first computer-adaptive tests comprised the major preoccupation of the language testers at the beginning of the 1980s. This was also the time when the first microcomputers were within reach for many applied linguists. Most of the papers in the early edited volumes in addition to journal articles (e.g., Larson and Madsen 1985) focused on issues associated with computer-adaptive testing. For example, reporting on a computer-adaptive test developed to increase efficiency of placement, Madsen (1991, p. 245) described the goal as follows: "intensive- English directors confirmed that the instrument they needed was an efficient and accurate ESL proficiency test rather than a diagnostic test." He describes the results of the research and development efforts in terms of the number of items required for placement, the mean number of items attempted by examinees, the mean amount of time it took students to complete the test, and students' affective responses to taking the test on the computer.

Other early developments appeared in a few papers exploring possibilities other than adaptivity, which were presented through the use of technology. The first issue of *Language Testing Update* at Lancaster University entitled "Innovations in language testing: Can the micro- computer help?" addressed the many capabilities of computers and how these might be put to use to improve language assessment for all test users, including learners (Alderson 1988). A paper in *CALICO Journal* at that time raised the need to reconcile the computer's capability for recording detailed diagnostic information with the test development concepts for proficiency testing, which are aimed to produce good total scores (Clark 1989). A few years later, Corbel (1993) published a research report at the National Centre for English Language Teaching and Research at Macquarie University, *Computer-Enhanced Language Assessment*, which also raised substantive questions about how technology might improve research and practice in language teaching and testing.

This early work expressed a vision of the potential significance of technology for changes and innovation in second language assessment, an agenda-setting collection of questions. However, the technology agenda for language assessment requires considerable infrastructure in addition to cross-disciplinary knowledge dedicated to problems in language assessment. At this time, decision-makers at the large testing companies, where such resources resided, apparently did not see technology-based assessment as a practical reality for operational testing programs. Instead, discussion of just a few innovative projects produced in higher education appeared (Marty 1981).

Significant advances involving computer recognition of examinees' constructed responses remained in research laboratories and out of reach for assessment practice

(Wresch 1993). This frustrating reality coupled with technical hardware and software challenges and the intellectual distance between most applied linguists and technology resulted in a slow start. By 1995, many applied linguists were voicing doubts and concerns about the idea of delivering high-stakes language tests by computer, fearing that the negative consequences would far outweigh any advantages. As it turned out, however, the technologies affecting language assessment did not wait for the approval and support of applied linguists. By the middle of the 1990s, many testing programs were beginning to develop and use computer-assisted language tests.

Major Contributions

The rocky beginning for technology in language assessment is probably forgotten history for most test users, as major contributions have now changed the assessment landscape considerably. Language test developers today at least consider the use of technology as they design new tests. Test takers and score users find online tests to be the norm like other aspects of language learning curricula and tools used in other facets of life. Contributions are complex and varied, but they might be summarized in terms of the way that technology has advanced language testing in four ways.

First, computer-adaptive testing has increased the efficiency of proficiency and placement testing. Many computer-adaptive testing projects have been reported regularly in edited books (i.e., Chalhoub-Deville 1999, and the ones cited earlier) and journal articles (e.g., Burston and Monville-Burston 1995). By evaluating examinees' responses immediately as they are entered, a computer-adaptive test avoids items that are either too easy or too difficult; such items waste time because they provide little information about the examinee's ability. In addition to creating efficient tests, these projects have raised important issues about the way language is measured, the need for independent items, and their selection through an adaptive algorithm. One line of research, for example, examines the effects of various schemes for adaptivity on learners' affect and test performance (Vispoel et al. 2000). Another seeks strategies for grouping items in a manner that preserves their context to allow several items to be selected together because they are associated with a single reading or listening passage. Eckes (2014), for example, investigated testlet effects in listening passages for a test of German as a foreign language.

Second, technology has prompted test developers to reconsider the constructs that they test. One example is the use of multimedia in testing listening comprehension. In the past, the testing of listening comprehension was limited to the examiner's oral presentation of linguistic input, either live or prerecorded, to a room full of examinees. Such test methods can be criticized for their failure to simulate listening as it occurs in many contexts, where visual cues are also relevant to interpretation of meaning. The use of multimedia provides test developers with the opportunity to contextualize aural language with images and to allow examinees to control their test-taking speed and requests for repetition. This option for construction of a test, however, brings interesting research questions about the nature of listening and the generalizability of listening across different listening tasks. Some of these questions are being explored in research on integrated tasks, which combine requirements for reading, writing, and speaking, for example. In this research, eye-tracking technology has proven useful for investigating how test takers interact with such tasks (Suvorov 2015).

Another example is the assessment of low-stakes dialogic speaking using Web cameras and videoconferencing software. Video simulates an interview in person with affordances for nonverbal skills, which are not available in monologic speech samples. For example, Kim and Craig (2012) found that linguistic performance on face-to-face English proficiency interviews was similar to performance on interviews conducted using videoconferencing software. The nonlinguistic cues such as gestures and facial expressions, however, were absent or difficult to see because of the small screen size. Advances in computer technology in research settings have made possible automatic assessment of dialogic oral interactions that include nonverbal communication. A computerized conversation coach developed at Massachusetts Institute of Technology, for example, provides summaries of oral and facial expressions such as head nodding and smiling through automated analysis in addition to speech recognition and prosody analysis in a simulated conversation (Hoque 2013). A third example is the use of actuators and sensors that sense changes in human emotion and mood, for instance, when a test taker is nervous during an oral interview. Although these technologies are not yet integrated in testing, Santos et al. (2016) are exploring the use of ambient intelligence to provide real-time natural interaction through visual, audio, and tactile feedback by a computer in response to changes in a learner affective state during a mock interview. These technological capabilities integrated into future assessments will allow test developers to assess both verbal and nonverbal aspects of speaking and in doing so will constantly require rethinking and investigating the construct meaning.

Third, natural language-processing technologies are being used for evaluating learners' spoken and written language. One of the most serious limitations with large-scale testing in the past was the over-reliance on selected-response items, such as multiple choice. Such items are used because they can be machine scored despite the fact that language assessment is typically better achieved if examinees produce language as they need to do in most language-use situations. Research on natural language processing for language assessment has recently yielded technologies that can score learners' constructed linguistic responses as well. A special issue of *Language Testing* in 2010 describes the research in this area and points to the use of these technologies in operational testing programs, typically for producing scores based on an evaluation of a response. Such evaluation systems are also being put to use for low-stakes evaluation and feedback for students' writing (Chapelle et al. 2015). Such work has advanced farther for responses that are written than those that are spoken.

Fourth, corpus linguistics is used to inform the design and validation of language assessments (Park 2014). A corpus can consist of texts produced by language learners or a collection of texts representing the target language-use domain relevant to score interpretation. Learner corpora are used by test developers to identify criterial linguistic features that appear in learners' language at particular stages of development. Such features can be used to produce descriptors for evaluating learners' constructed responses or to investigate the language elicited from particular test tasks.

Corpora representing the target language-use domain can be used to identify lexical, structural, and functional content that characterizes a particular language domain. One purpose of defining the domain is to ensure that test tasks are modeled on tasks that test takers will perform in the target domain (e.g., Biber 2006). Such an investigation can result in selection of specific linguistic features for test items as Voss (2012) did by sampling collocations from a corpus of academic language. In this case, the corpus was also used to verify frequent and possible collocations to inform a partial-credit scoring procedure. Similarly, reading and listening passages can be selected or developed with appropriate difficulty levels based on the frequency of lexis in the passage aligned with characteristics identified in corresponding proficiency levels. Using frequency and sentence length data, for example, standardized Lexile® scores for reading passages are used to complement assessment results with level-appropriate instruction and reading ability levels (Metametrics 2009). The systematicity and empirical basis of linguistic analysis during test development are an important part of the evidence in a validity argument for the test score interpretations.

These technical advances in test methods need to be seen within the social and political contexts that make technology accessible and viable to test developers, test takers, and test users. Not long ago most test developers felt that the operational constraints of delivering language tests by computer may be insurmountable. Today, however, many large testing organizations are taking advantage of technical capabilities that researchers have been investigating for at least the last 20 years. As computer-assisted language assessment has become a reality, test takers have needed to reorient their test preparation practices to help them prepare.

Work in Progress

The primary impetus for using technology in language assessment was for many years to improve the efficiency of testing practices and thus much of the work in progress has centered on this objective. Research is therefore conducted when testing practices are targeted for replacement by computer-assisted testing for any number of reasons such as an external mandate. The objective for research in these cases is to demonstrate the equivalence of the computer-assisted tests to the existing paper-and-pencil tests. For example, such a study of the Test of English Proficiency developed by Seoul National University examined the comparability of computer-based and paper-based language tests (CBLT and PBLT, respectively). Choi et al. (2003) explained the need for assessing comparability in practical terms: "Since the CBLT/CALT version of the [Test of English Proficiency] TEPS will be used with its PBLT version for the time being, comparability between PBLT and CBLT is crucial if item statistics and normative tables constructed from PBLT are to be directly transported for use in CBLT" (Choi et al. 2003, p. 296). The study, which used multiple forms of analysis to assess comparability of the constructs measured by the two tests, found support for similarity of constructs across the two sets of tests, with the listening and grammar sections showing the strongest similarities and the reading sections showing the weakest.

In addition to the practical motivation for assessing similarity to determine whether test scores can be interpreted as equivalent, there is an important scientific question to be investigated as well: what important construct-relevant differences in language performance are sampled when technology is used for test delivery and response evaluation. Unfortunately, few studies have tackled this question (Sawaki 2001). The use of technology for test delivery is frequently a decision that is made before research, and therefore the issue for practice is how to prepare the examinees sufficiently so that they will not be at a disadvantage due to lack of computer experience. For example, Taylor et al. (1999) gave the examinees a tutorial to prepare them for the computer-delivered items before they investigated the comparability of the computer-based and the paper-and-pencil versions of test items for the (TOEFL). In this case, the research objective is to demonstrate how any potential experience-related difference among test takers can be minimized. The need for tutorials is disappearing as younger learners grow up with computer technology. Computer and language literacy develop together as the use of touch-screen tablets in homes and early education is increasing (Neumann 2016). In response such new practices for literacy development, The Cambridge English Language Assessment allows young test takers to choose their preferred mode of test delivery by taking the test on a computer or on paper (Papp and Walczak 2016). The results of research investigating performance on both show that the two delivery modes were comparable, that "children are very capable of using computers, and that they especially like using iPads/tablets" (p. 168).

As technology has become commonplace in language education, researchers and developers hope to expand the uses and usefulness of language tests through innovation. For example, the DIALANG project, an an Internet-based test, developed shortly after the advent of the Web (Alderson 2005), was intended to offer diagnostic information to learners to increase their understanding of their language learning. Whereas DIALANG was intended to have extensive impact on language learners due to its accessibility on the Web, other assessments aimed at learning appear in computer-assisted language learning materials. Longman English Interactive (Rost 2003), for example, includes assessments regularly throughout the process of instruction to inform learners about how well they have learned what was taught in each unit. Such assessments, which also appear in many teacher-made materials, use technology to change the dynamic between test takers and tests by providing learners a means for finding out how they are doing, what they need to review, and whether they are justified in their level of confidence about their knowledge. These same ideas about making assessment available to learners through the delivery of low-stakes assessment are migrating to the next generation of technologies.

For example, Palomo-Duarte et al. (2014) describe a low-stakes test of vocabulary that learners can take on their smartphones by downloading an app. Also, designed to meet student demand, many apps have been created to accompany language learning or as practice test for standardized language tests such as TOEFL and IELTS.

For high-stakes testing, in contrast, lack of adequate security poses a thorny problem for assessment on mobile platforms. However, because mobile devices with multimedia capabilities and Internet access are becoming so commonplace, the development of low-cost, large-scale, high-stakes language tests with multimodal interaction is enticing. For example, two universities in Spain are exploring the delivery of the Spanish University Entrance Examination on mobile devices (García Laborda et al. 2014). The mobile-enhanced delivery of the Spanish test includes assessment of grammar, reading, writing, listening, and speaking with a combination of (automated rating and responses assessed later by human raters). Currently, such devices are best suited for listening and speaking tasks because small screen sizes on mobile phones make appropriate reading tasks difficult to construct. Technological limitations also affect the expected written responses that can be requested of test takers. Producing written language on a smartphone entails a number of fundamental differences from writing at a keyboard, and therefore, the device needs to be considered carefully in the design of test tasks. Smartphone testing issues are undoubtedly entering into mainstream language testing because their reach extends even beyond that of the Internet. In physical locations where the Internet connection is slow or nonexistent, language tests have been administered using the voice and SMS texting technologies of mobile phones (Valk et al. 2010). Delivery of assessments to students in remote areas is possible with these platforms even if supplemental paper-based materials are necessary.

All of this language testing development relies on significant software infrastructure, and therefore another area of current work is the development of authoring systems. Due to limitations in the existing authoring tools for instruction and assessment, most language-testing researchers would like to have authoring tools intended to address their testing goals directly, including the integration of testing with instruction, analysis of learners' constructed responses, and capture and analysis of oral language. As such, capabilities are contemplated for authoring tools, as are new ways for conceptualizing the assessment process. Widely used psychometric theory and tools were developed around the use of dichotomously scored items that are intended to add up to measure a unitary construct. The conception of Almond et al. (2002) underlying their test authoring tools reframes measurement as a process of gathering evidence (consisting of test takers' performance) to make inferences about their knowledge and capabilities. The nature of the evidence can be, but does not have to be, dichotomously scored items; it can also be the results from a computational analysis of learners' production. Inferences can be made about multiple forms of knowledge or performance. The emphasis on evidence and inference underlies plans for developing authoring tools for computer-assisted testing that can include a variety of types of items and can perform analysis on the results that are obtained – all within one system.

Problems and Difficulties

With the intriguing potentials apparent in current work, many challenges remain, particularly in view of the changing technologies. Testing programs need to have built-in mechanisms for updating software, hardware, and technical knowledge of employees. Large testing companies with the most resources may be the most able to keep up with changes. To some extent they have done so by increasing fees for those using their tests. In some cases costs are borne by language programs, but in many other cases, the costs are passed on to those who are least able to pay – the test takers themselves. Small testing organizations, publishing companies for whom testing is just one part of their overall profile, as well as school-based testing programs have to rely on strategic partnerships to combine expertise, limited resources, and technologies. Navigation of these waters in a quickly changing environment requires exceptionally knowledgeable leadership.

Challenges that may be less evident to test users are those that language-testing researchers grapple with as they attempt to develop appropriate tests and justify their use for particular purposes. As Bachman (2000, p. 9) put it, “the new task formats and modes of presentation that multimedia computer-based test administration makes possible raise all of the familiar validity issues, and may require us to redefine the very constructs we believe we are assessing.” For example, Chapelle (2003) noted that in a computer-assisted reading test, the test tasks might allow the test takers access to a dictionary and other reading aids such as images. In this case, the construct tested would be the ability to read with strategic use of online help. The reading strategies entailed in such tasks are different from those used to read when no help is available, and therefore the definition of strategic competence becomes critical for the construct assessed. Should test takers be given access to help while reading on a reading test? One approach to the dilemma is for the test developer to decide whether or not access to help constitutes an authentic task for the reader. In other words, if examinees will be reading online with access to help, such options should be provided in the test as well. However, the range of reading tasks the examinees are likely to engage in is sufficiently large and diverse to make the authentic task approach unsatisfactory for most test uses. The reading construct needs to be defined as inclusive of particular strategic competencies that are required for successful reading across a variety of contexts.

A second example of how technology intersects with construct definition comes from tests that use natural language processing to conduct detailed analyses of learners’ language. Such analyses might be used to calculate a precise score about learners’ knowledge or to tabulate information about categories of linguistic knowledge for diagnosis. In either case, if an analysis program is to make use of such information, the constructs assessed need to be defined in detail. A general construct definition such as “speaking ability” does not give any guidance concerning which errors and types of disfluencies should be considered more serious than others, or which ones should be tabulated and placed in a diagnostic profile. Current trends in scoring holistically for overall communicative effectiveness circumvent the need for taking a close linguistic look at constructed responses. One of the few studies to grapple with this issue (Coniam 1996) pointed out the precision afforded by the computational analysis of the learners' responses far exceeded that of the construct of listening that the dictation test was measuring. To this point assessment research has not benefited from the interest that second language acquisition researchers have in assessing detailed linguistic knowledge; it remains a challenge (Alderson 2005).

Another challenge that faces language-testing researchers is the need to evaluate computer-assisted language tests. As described earlier, current practices have focused on efficiency and comparability. However, one might argue that the complexity inherent in new forms of computer-assisted language assessment should prompt the use of more sensitive methods for investigating validity. When the goal of test development is to construct a more efficient test, then efficiency should clearly be part of the evaluation, but what about computer-assisted tests that are intended to provide more precise measurement, better feedback to learners, or greater accessibility to learners? If the scores obtained through the use of natural language-processing analysis are evaluated by correlating them with scores obtained by human raters or scores obtained with dichotomously scored items (e.g., Henning et al. 1993), how is the potential additional value of the computer to be detected?

In arguing for evaluation methods geared toward computer-assisted language tests, some language-testing researchers have focused on interface issues (Fulcher 2003) – an important distinction for computer-assisted tests. It seems that the challenge is to place these interface issues within a broader perspective on validation that is not overly preoccupied by efficiency and comparability with paper-and-pencil tests. Chapelle et al. (2003), for example, frame their evaluation of a Web-based test in broader terms, looking at a range of test qualities. Chapelle and Douglas (2006) suggest the continued need to integrate the specific technology concerns into an overall agenda for conceptualizing validation in language assessment that includes the consequences of test use. Technology reemphasizes the need for researchers to investigate the consequences of testing. Such consequences might include benefits such as raising awareness of the options for learning through technology.

Future Directions

These two sets of challenges – the obvious ones pertaining to infrastructure and the more subtle conceptual issues evident to language-testing researchers – combine to create a third issue for the field of applied linguistics. How can improved knowledge about the use of technology be produced and disseminated within the profession? What is the knowledge and experience that graduate students in applied linguistics should attain if they are to contribute to the next generations of computer-assisted language tests? At present, it is possible to identify some of the issues raised through the use of technology that might be covered in graduate education, but if graduate students are to dig into the language-testing issues, they need to be able to create and experiment with computer-based tests.

Such experimentation requires authoring tools that are sufficiently easy to learn and transportable beyond graduate school. Commercial authoring tools that are widely accessible are not particularly suited to the unique demands of language assessment such as the need for linked items, the evaluation of learners' oral and written production, and the collection of spoken responses. As a consequence, many students studying language assessment have no experience in considering the unique issues that these computer capabilities present to language testing. In a sense, the software tools available constrain thinking about language assessment making progress evolutionary rather than revolutionary (Chapelle and Douglas 2006).

More revolutionary changes will probably require graduate students educated in language testing in addition to other areas of applied linguistics. For example, students need to be educated in corpus linguistics to conduct appropriate domain analyses as a basis for test development (e.g., Biber 2006). Education in second-language acquisition is needed too for students to use learner corpora for defining levels of linguistic competence (Saville and Hakey 2010). Education in world Englishes is needed to approach issues of language standards (Mauranen 2010). These and other aspects of applied linguistics appear to be critical for helping to increase the usefulness of assessment throughout the educational process, strengthen applied linguists' understanding of language proficiency, and expand their agendas for test validation.

Cross-References

- ▶ Cognitive Aspects of Language Assessment
- ▶ Using Portfolios for Assessment/Alternative Assessment
- ▶ Utilizing Accommodations in Assessment

Related Articles in the Encyclopedia of Language and Education

Rémi A. van Compernolle: Sociocultural Approaches to Technology Use in Language Education. In Volume: Language, Education and Technology

Kevin M. Leander, Cynthis Lewis: Literacy and Internet Technologies. In Volume: Literacies and Language Education

John Thorne: Technologies, Language and Education: Introduction. In Volume: Language, Education and Technology

Paula Winke, Daniel R. Isbell: Computer-Assisted Language Assessment. In Volume: Language, Education and Technology

References

Alderson, J. C. (1988). *Innovations in language testing: Can the microcomputer help? Special Report No 1 Language Testing Update*. Lancaster: University of Lancaster.

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

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning and Assessment, 1*(5). Available from http://www.jtla.org

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1–42.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Burstein, J., Frase, L., Ginther, A., & Grant, L. (1996). Technologies for language assessment. *Annual Review of Applied Linguistics, 16*, 240–260.

Burston, J., & Monville-Burston, M. (1995). Practical design and implementation considerations of a computer-adaptive foreign language test: The Monash/Melbourne French CAT. *CALICO Journal, 13*(1), 26–46.

Chalhoub-Deville, M. (Ed.). (1999). *Development and research in computer adaptive language testing*. Cambridge: University of Cambridge Examinations Syndicate/Cambridge University Press.

Chapelle, C. A. (2003). *English language learning and technology: Lectures on applied linguistics in the age of information and communication technology.* Amsterdam: John Benjamins Publishing.

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

Chapelle, C., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20*(4), 409–439.

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Diagnostic assessment with automated writing evaluation: A look at validity arguments for new classroom assessments. *Language Testing, 32*(3), 385–405.

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*(3), 295–320.

Clark, J. L. D. (1989). Multipurpose language tests: Is a conceptual and operational synthesis possible? In J. E. Alatis (Ed.), *Georgetown university round table on language and linguistics. Language teaching, testing, and technology: Lessons from the past with a view toward the future* (pp. 206–215). Washington, DC: Georgetown University Press.

Coniam, D. (1996). Computerized dictation for assessing listening proficiency. *CALICO Journal, 13*(2–3), 73–85.

Corbel, C. (1993). Computer-enhanced language assessment. In G. Brindley (Ed.), *Research report series 2, National Centre for English Language Teaching and Research*. Sydney: Marquarie University.

Dunkel, P. (Ed.). (1991). *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39–61.

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

García Laborda, J. G., Magal-Royo, T. M., Litzler, M. F., & Giménez López, J. L. G. (2014). Mobile phones for Spain's university entrance examination language test. *Educational Technology & Society, 17*(2), 17–30.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* London: Sage.

Henning, G., Anbar, M., Helm, C., & D'Arcy, S. (1993). Computer-assisted testing of reading comprehension: Comparisons among multiple-choice and open-ended scoring methods. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research*. Alexandria: TESOL.

Hoque, M. E. (2013). *Computers to help with conversations: Affective framework to enhance human nonverbal skills* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 0830325).

Kim, J., & Craig, D. A. (2012). Validation of a videoconferenced speaking test. *Computer Assisted Language Learning, 52*(3), 257–275.

Larson, J. W., & Madsen, H. S. (1985). Computer-adaptive language testing: Moving beyond computer-assisted testing. *CALICO Journal, 2*(3), 32–36.

Madsen, H. S. (1991). Computer-adaptive test of listening and reading comprehension: The Brigham Young University approach. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 237–257). New York: Newbury House.

Marty, F. (1981). Reflections on the use of computers in second language acquisition. *Studies in Language Learning, 3*(1), 25–53.

Mauranen, A. (2010). Features of English as a lingua franca in academia. *Helsinki English Studies, 6*, 6–28.

Metametrics. (2009). The Lexile framework for reading. Retrieved from http://www.lexile.com

Neumann, M. M. (2016). Young children's use of touch screen tablets for writing and reading at home: Relationships with emergent literacy. *Computers & Education, 97*, 61–68.

Palomo-Duarte, M., Berns, A., Dodero, J. M., & Cejas, A. (2014). Foreign language learning using a gamificated APP to support peer-assessment. In *Proceedings of TEEM' 14: Second international conference on technological ecosystem for enhancing multiculturality*, Salamanca, Volume 1.

Papp, S., & Walczak, A. (2016). The development and validation of a computer-based test of English for young learners: Cambridge English young learners. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (Educational linguistics, Vol. 25, pp. 139–190). New York: Springer.

Park, K. (2014). Corpora and language assessment: The state of the art. *Language Assessment Quarterly, 11*(1), 27–44.

Rost, M. (2003). *Longman English interactive*. New York: Pearson Education.

Santos, O. C., Saneiro, M., Boticario, J. G., & Rodriquez-Sanchez, M. C. (2016). Toward interactive context-aware affective educational recommendations in computer- assisted language learning. *New Review of Hypermedia and Multimedia, 22*(1), 27–57.

Saville, N., & Hakey, R. (2010). The English language profile – The first three years. *English Language Journal, 1*(1), 1–14.

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

Stansfield, C. (Ed.). (1986). *Technology and language testing*. Washington, DC: TESOL Publications.

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing, 32*(4), 463–483.

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*(2), 219–274.

Valk, J. H., Rashid, A. T., & Elder, L. (2010). Using mobile phones to improve educational outcomes: An analysis of evidence from Asia. *The International Review of Research in Open and Distance Learning, 11*(1), 117–140.

Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement, 37*(1), 21–38.

Voss, E. (2012). *A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3539432).

Wresch, W. (1993). The imminence of grading essays by computer – 25 years later. *Computers and Composition, 10*(2), 45–58.