# 20 YEARS OF TECHNOLOGY AND LANGUAGE ASSESSMENT IN LANGUAGE LEARNING & TECHNOLOGY

**Carol A. Chapelle, Iowa State University**
**Erik Voss, Northeastern University**

This review article provides an analysis of the research from the last two decades on the theme of technology and second language assessment. Based on an examination of the assessment scholarship published in *Language Learning & Technology* since its launch in 1997, we analyzed the review articles, research articles, book reviews, and commentaries as developing one of two primary thrusts of research on technology and language assessment: technology for efficiency and technology for innovation.

**Keywords:** Testing

**APA Citation:** Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in Language Learning & Technology. *Language Learning & Technology*, *20*(2), 116–128. Retrieved from http://llt.msu.edu/issues/june2016/chapellevoss.pdf

**Received:** June 2, 2015; **Accepted:** August 17, 2015; **Published:** June 1, 2016



## INTRODUCTION

Assessment of learners' language ability is an important part of language education, which has been affected by computer technology at least as significantly as language learning has. Because of the significance of language assessment for language teachers, software developers, applied linguists, and learners, articles in *Language Learning & Technology* (*LLT*) have contributed to chronicling the developments in language assessment technologies. Throughout these papers, the terms language testing and language assessment are used to denote the process of systematically gathering data from learners to make interpretations about their language abilities and decisions about their future. These processes of making interpretations and using those interpretations for decision-making have been taking place in language education for a long time, predating the arrival of computer technology. However, the new technologies appearing in the second half of the 20th century have been applied to these processes in hopes of improving them. In some cases, such improvements appear in the form of efficiencies in existing practices, but another important thrust of research in this area is the innovation afforded by technology for language testing. In this paper we provide an overview of the ways in which language assessment has been presented over 20 years of *LLT* and how this presentation reflects the broader academic scholarship on the use of technology in language testing. This review shows that the main themes in this area have been developed in a manner that reflects the larger field, beginning with the excitement offered by the psychometric advances that gave birth to computer-adaptive testing, and developing into a range of innovative assessments and uses that draw on many areas of applied linguistics. We begin with a brief description of our approach to looking back at *LLT*. We then summarize the key threads that have appeared in *LLT* and connect them to the larger fabric of language testing research.

## LOOKING BACK AT LLT

*LLT* has published many articles on language assessment throughout its 20 years. We therefore approached our review by looking back at all of the issues from the perspective of language testers. Specifically, we searched articles published in *LLT* for the terms *test*, *testing*, *assessment*, *evaluation*, *validity*, *validation*, *validating*, *usefulness*, and *argument* using the search field on the journal's website, which returned papers published from 1997 to 2015. All results were listed in a Word document omitting

any duplicates, announcements, news, reference lists, and brief introductions from editors. The 198 documents included papers on learning, instruction, and research, all of which contained sections on assessment. The papers were examined manually to assess whether or not the primary focus of each paper was assessment, and if it was not, it was not included in the sample that we examined for this review article. A total of 25 documents met our inclusion criteria. The final list contained 8 articles reporting empirical research in addition to 9 other review or theoretical articles, for a total of 17 articles. We also included 4 book reviews, 3 software reviews, and a commentary, creating a set of 25 papers representing all of the genres that have been published in *LLT*.

Based on our analysis of each document, we categorized its contribution to a particular aspect of the overall theme of technology and second language assessment. The categorization was done based on our knowledge of the work in the area of second language assessment, which includes research encompassing a range of assessments and assessment uses. The topics we identified appear in a roughly chronological progression from issues of efficiency including automation of existing practices in place to more recent innovations in language assessment.

## TECHNOLOGY FOR EFFICIENCY

Technology in language testing, like in language teaching, was introduced in the 1960s with the motivation of making the testing process more efficient. García Laborda stated the efficiency case clearly in his 2007 review article in which he concluded, "the benefits of online testing should overcome any of its drawbacks, as it can be faster, more efficient, and less costly than traditional paper-and-pencil testing. Additionally, multimedia prompts can help make the test feel more 'real.' Adaptive tests can facilitate the difficult task of rapid diagnosis, and self-correcting tests can accelerate the process of correction, feedback, and reporting" (García Laborda, 2007, p. 8).

The excitement for the efficiency movement expressed in García Laborda's (2007) paper began primarily with language testers working with high-stakes testing. In large-scale, high-stakes testing, new efficiencies could have significant financial benefits for testing organizations that were prepared to build the hardware and software infrastructure as well as the knowledge base to take advantage of the potential. Chronologically, the first big development was computer-adaptive testing, which provided a means of tailoring a test to each student interactively during test taking. Efficiency was also evident from the first uses of automated writing evaluation (AWE), whose promise was, and in part remains, the use of technology to perform at least some of the time intensive work of evaluating students' essays. Computer-adaptive testing and AWE are two themes developed in the papers in *LLT*. A third general theme is one that motivates efficiency-oriented research on all computer-assisted language tests, that is the concern for comparison of score meaning for computer-assisted tests and those delivered through other mechanisms.

### Adaptive Testing

Computer-adaptive language testing refers to the research and practice that goes into development and validation of tests that use technology to interactively monitor test takers' performance and branch, or adapt, based on an algorithm specified by the test developer. As professionals in language learning and technology know, branching can be developed to accomplish many different purposes, and the same is true in language testing. But early computer-adaptive language testing referred to a particular type of algorithm that relies on a psychometric method called item response theory (IRT) to control the adaptivity based on test takers' performance on each item on the test. This type of computer adaptive language testing was the primary focus of the papers in *LLT*.

Between the years 1997 to 2001, articles about L2 adaptive testing in *LLT* included two review articles, two book reviews, and one commentary on another *LLT* article. In the first issue of *LLT*, Brown's (1997) review article described multiple issues of efficiency such as test production and delivery (e.g., item banking), but primary attention is given to the efficiencies to be gained by computer-adaptive language

testing. Dunkel (1999) expanded on many of the issues Brown's paper introduced in her review article which provided an introduction to computer-adaptive testing aimed at prospective test developers. She reviewed the advantages of computer-adaptive testing including some of the same advantages that have been identified for computer-assisted language learning: each student can work at his or her own pace, the adaptivity results in each student being presented with tasks appropriate to his or her level, students can receive immediate feedback on the correctness of responses, and multimedia presentation can depict authentic situations of language use.

The affordances of computer-assisted language learning have additional meaning in a testing context. The monitoring of students' work that allows for adaptivity also creates the opportunity to gather an additional type of data that can be relevant to assessing students' abilities: the amount of time they spend on each item can be recorded and used as an indicator of the automaticity or fluency of test takers' performance. Adaptivity also means that testing time is used wisely by not having students respond to test items that are too easy or too difficult. Time spent on items poorly suited to the test takers' ability level typically provide no measurement information. The provision of feedback to students also means that taking the test may provide an opportunity for students' learning. The adaptive generation of a test tailored to each student means that test security is strengthened: each student takes a different version of the test, making traditional cheating ineffective. The use of multimedia in testing allows language testers to assess language skills that may be different than those assessed by audio alone. To realize these potentials for language testing, Dunkel (1999) laid out the numerous technical issues that need to be considered in the design and development of a computer-adaptive test, some of which are the same as one would recognize for any language test design. However, computer-adaptive testing presents some new demands for language testing, as well.

Dunkel's (1999) article is a good entry point to the topic for readers of two books on computer-adaptive testing that were reviewed in *LLT. Computerized adaptive testing: A primer (2nd ed.)* edited by Wainer (2000) was reviewed by Norris (2001a). As Norris pointed out, this book presents and discusses foundational issues for the development, use, and evaluation of all computer-adaptive testing, not limited to language testing. The review of the book also positions computer adaptive tests as one type of computer-based test and highlights the need to "evaluate the extent to which CATs are appropriate for the kinds of inferences and purposes we need to address in language assessment" (Norris, 2001a, p. 26). The other book on computer-adaptive testing was *Issues in computer-adaptive testing of reading proficiency*, edited by Chalhoub-Deville (1999). Reviewer Fernández-García (2001) introduces the volume as a collection of papers, many of which come from an invitational conference focused on the range of issues from the technical to conceptual in the assessment of construct of L2 reading. The volume is notable for its inclusion of disparate voices describing the perspectives from various disciplines that pertain to computer-adaptive language testing design and development. The contributions reveal the complexity and range of the issues involved in computer adaptive testing of L2 reading and raise significant challenges.

Identifying challenges was also the goal of a commentary by Norris (2001b) on the article "Comparing Examinee Attitudes Toward Computer-Assisted and Other Oral Proficiency Assessments" by Kenyon and Malabonga (2001). Their article described an alternative to the IRT-driven, item-based adaptivity that is highlighted in the other articles. Because an oral proficiency assessment is a speaking test, computational scoring of students' responses was seen as impossible, and therefore the use of the computer program to control adaptive decision-making was not plausible. Instead, the algorithm used a combination of students' judgments about the difficulty of items as they completed them and a rule for intermittently selecting more difficult ones than the students' choices would suggest. The test actually began by asking students to judge their own levels. The commentary by Norris (2001b) pointed out this different interpretation of the meaning of adaptivity and raised an array of questions about assessing productive language performance such as speaking ability with contemporary adaptive testing technology. Both the original article and the commentary were predictive of the types of discussions that have been taking

place over the past years, as adaptivity is understood more broadly than it once was (e.g., Mislevy, Chapelle, Chung, & Xu, 2007).

**Automated Writing Evaluation**

Automated writing evaluation (AWE), also known as automated essay evaluation, was introduced in the 1960s with the promise of possibly improving efficiency in language teaching and testing by automating the time-consuming work of evaluating students' writing in English. The initial excitement for this prospect did not materialize into useable technologies right away, however. The development of systems to achieve high quality automated essay evaluation has become a complex cross-disciplinary research issue for the field of language assessment. The past and current research on the problem of automating the evaluation process for students' writing is covered in the *Handbook of Automated Essay Evaluation: Current Applications and New Directions* edited by Shermis and Burstein (2013). Their book, reviewed by Zhang (2014), highlights major issues related to writing instruction using AWE, existing AWE systems, constructs that AWE systems assess, as well as issues of reliability and validity.

**Assessing Comparability of the Old and New**

A central concern for language testers attempting to increase efficiencies in testing is the comparability of the new version of a test with an existing test. The assumption behind this way of thinking is that existing practices serve as a gold standard, or at least a known standard, against which new test versions can be measured. In the case of traditional IRT-based adaptivity, test developers can point to test characteristics such as shorter testing time, fewer items, and higher reliability—or at least, a more accurate estimate of reliability—for each test score obtained from the computer-adaptive test than from its paper-and-pencil counterpart. Comparisons are also drawn between the effects of each test version on test taker affect. All of the specific comparisons seek to demonstrate the extent to which the scores on one version of the test can be considered to be equivalent to scores on the other version. In 2001, one review article and two research articles addressed the issue of comparability between conventional tests, technology-mediated tests, and computer-based tests in reading and speaking assessments (Kenyon & Malabonga, 2001; Norris, 2001b, Sawaki, 2001), and then, in 2004, another research article looked at comparability on a writing test (Wolfe & Manalo, 2004).

The issue of comparability between computerized and conventional L2 reading was addressed in the review article "Comparability of conventional and computerized tests of reading in a second language" (Sawaki, 2001). Sawaki reviewed the existing knowledge base pertaining to the question of whether or not test scores from computer-assisted reading tests and their paper-and-pencil counterparts should be expected to be equivalent. The domain of her review included research in measurement about comparability of computer-assisted tests and paper-and-pencil tests across a number of different content areas. The conclusion was that "the empirical findings as to comparability of conventional and computerized tests are rather mixed" (p. 44). But exactly what is *mixed*? How does one compare the two modes of testing? Sawaki's review answers this question by describing the various comparisons that need to be made including the comparability of task content and administration conditions across modes of presentation, the psychometric criterion of stability of item parameter estimates, the plausibility of linking tests across modes, the potential interaction of examinee characteristics and testing conditions, the comparability of decisions, and the impact of the introduction of computerized tests to examinees. These are all aspects of the test and testing process that come into play when developing a validity argument for test score interpretation and use. Sawaki also probed the nature of the construct of reading ability and how it may be affected by the mode of text a person reads by examining the research in ergonomics, education, psychology, and L1 reading. She concluded, "the general trends found in these studies indicate that comprehension of computer-presented texts is, at best, as good as that of printed texts, and that reading speed may or may not be affected by mode of presentation" (p. 49). She included in her interpretation of these findings a pertinent sociological factor: that the issue of computer familiarity that seemed central to

comparability questions in the past was, in 2001, no longer so important. In 2001, it would have been unusual to find readers unfamiliar with reading on a screen, and today, one might even ask if some test takers would be more familiar with reading on a screen than they would be reading text on paper.

Even if people are accustomed to reading a computer screen, they may be less comfortable talking to a computer screen. In a research article, Kenyon and Malabonga (2001) evaluated attitudes and perceptions from test takers after completing a speaking assessment with two technology-mediated tests, a Simulated Oral Proficiency Interview, which is administered using a tape recorder, and a new Computerized Oral Proficiency Instrument, which has features of an adaptive test as described above. The students in the Spanish group of the participants also completed a face-to-face Oral Proficiency Interview. The computerized version allowed some measure of adaptivity with students' judgments playing a role in task selection. The authors explained that the goal of the adaptivity was to help students to do their best on the test. This goal would be undermined if students were uncomfortable using the technology. With this as the primary concern about the adaptivity, the research focused on the potential "interaction of examinee characteristics and testing conditions" (Sawaki, 2001, p. 42), one of the areas of comparison Sawaki outlined in her review. As Norris (2001b) pointed out in his commentary of this study, other areas of comparison remained to be done, some of which extend beyond comparison studies. However, the manner in which Kenyon and Malabonga (2001) conceptualized their study—focusing on the area of greatest concern—is a typical strategy in validation research and demonstrates the reason that validation research is an ongoing process.

Another empirical study compared the use of the computer for administering a writing test relative to use of the more traditional paper-based writing test. The paper by Wolfe and Manalo (2004) compared handwritten essays and essays written using a word processor for the written section of the Test of English as a Foreign Language. Test takers were given a choice of medium. Results indicated differences among proficiency level as well as geographic region, native language, gender, and age.

Throughout the efficiency-oriented papers, it is evident that the primary concern for language testers is to improve the testing process by making it faster, more reliable, and more efficient. In her commentary, Chalhoub-Deville (2001) discusses computer-adaptive testing highlighting its advantages, but at the same time, being critical of efficiency as a sole direction for the developments in technology and language assessment. Instead, she suggests, "advances in technology should encourage test developers to move beyond the thinking that has long dominated paper-and-pencil testing and inspire the use of 'disruptive' applications, by which assessments are conceptualized and implemented in innovatively different ways" (p. 97). In other words, traditional computer-adaptive testing with item-level adaptivity can be considered a constraining force in view of the possibilities. We could add that a research program dominated by concerns about the comparability of the old with the new is missing opportunities for innovation.

## TECHNOLGOY FOR INNOVATION

An innovative agenda for language assessment extends beyond the goal of making more efficient tests to expanding the uses of assessment and their usefulness. Roever's (2001) review article in *LLT* connected the potential for innovation with the introduction of language testing on the web, pointing out that the affordances of the web—particularly its accessibility for prospective authors, teachers, and students—increase the opportunities for assessment to be integrated into student learning. The access to the material capacity for using technology is clearly essential for innovation to emerge, but it also requires innovators with sufficient understanding and motivation to think beyond efficiency. Innovation is a way of thinking about language assessment that emerges when language testers, teachers, and students consider technology as a resource for improving methods and increasing uses (Chapelle & Douglas, 2006). Successful innovation can also result in new knowledge about the intersection of technology with assessment. A number of issues in innovation have been raised in *LLT* over the past years including the use of assessment to increase and improve opportunities for learning, the importance of rethinking the

language constructs to be measured, the need to investigate potential impacts of innovations, and the engagement of the profession in new language assessments through access to authoring tools for assessment.

**Opportunities for Learning**

A central idea in the work on innovation in language assessment is that test takers should actually be given opportunities to learn from both the process and the results of test taking. The unique capabilities of technology are ideally suited to play a role in this vision because of their capacity for individual treatment of test takers as learners, their natural place in distance learning programs, and their potential for expanding learning processes.

*LLT* has published several papers showing examples of some of the capacities technology provides for individualized analysis of learners' language, feedback, and reporting. These characteristics of assessment in support of learning are evident in the descriptions of AWE systems. In 2008, Godwin-Jones reviewed web-based resources for online writing including informal and formal online writing assessment. The review provided an introduction to products that evaluate surface features, such as spelling and grammar in addition to those that perform automated scoring and offer students individualized feedback on their writing. One challenge, however, is to develop proofing tools that are sophisticated enough to provide useful feedback to different levels of English language learners. A second challenge that Godwin-Jones highlighted is that many products focus on specific sentence level grammar and vocabulary usage without addressing more general issues of global composition. A closer look at one system appeared four years later in review of Criterion, the AWE system developed initially for high-stakes testing (Lim & Kahng, 2012). The underlying software architecture of linguistic feature detection is put to work to identify errors in written text. The results of the error identification are used by the system to generate feedback to writers. As Lim and Kahng explained, this detailed information about students' performance can also be summarized in reports for the teacher. Their review of Criterion, which encompasses its use as a tool for both testing and learning, demonstrates how software development plays a central role in pivoting attention to learning.

*LLT*'s focus on software in the reviews was complemented by an article reporting the results of a study investigating the use of an AWE system. Chen and Cheng (2008) evaluated the use of MyAccess! software in three EFL college writing classrooms for 6–16 weeks. Each course implemented the software in a different way. The authors were interested in how students and instructors adapted to a pedagogical tool that provides immediate computer-generated scores along with diagnostic feedback. The "diagnostic feedback function [seemed] pedagogically appealing for formative learning" (p. 97). They found, however, that the effectiveness of the software depended on students' familiarity with it and their ability to use it. For example, a course that used only AWE for feedback was frustrating to students and limited their perception of the writing development process. Students preferred a combination of automated scoring on early drafts followed by human feedback later in the writing process. This combination of scores and diagnostic feedback from the system along with human guidance and feedback based on a sound pedagogical foundation shows the most promise to support the assessment of and for learning.

Technology can also be integral to learning when it is used in distance learning programs delivered via the web. *LLT* published an article describing research evaluating the effectiveness of content delivered through "CD-ROM/DVD programs, online content-based web pages, and synchronous bimodal chat that includes sound and text" in hybrid and distance-learning (DL) college Spanish language courses (Blake et al., 2008). Selecting appropriate language assessments for online instruction is just as important as for face-to-face courses. Blake et al. (2008) elected to use a phone delivered automated speaking assessment to measure oral proficiency at the end of the course as evidence of learning in an online context. Based on the results of the speaking assessment, the authors claim that oral proficiency of students in DL formats was similar to that of classroom learners and that the online speaking assessment was "capable of

distinguishing different levels of oral proficiency that roughly correspond to first-year, second-year, and heritage students" (p. 123).

A third area of important innovation focuses on the use of technology to construct assessments that expand the possibilities for student learning beyond what is possible in a traditional classroom. The bounds for innovation and language assessment in this area are unknown, but *LLT* has some examples of ideas with the potential to change learning. An article by Teo (2012) described computer-assisted dynamic assessment to promote metacognitive reading strategies in a Freshman English class in Taiwan. Teo's (2012) article explained that assessment can be difficult in large language classes, particularly those providing human-to-human mediation for dynamic assessment. The computer attempted to fill the role of a human mediator as students interacted and responded to mediated feedback in a web-based computerized dynamic assessment (C-DA). "The C-DA program consisted of mediation that was designed to improve the learners' metacognitive strategies, especially with the intention of training them to be strategic and reflective readers" (Teo, 2012, p. 12). Data about students' performance were captured by the computer for the instructor to review. Teo's action research project explained how software could create opportunities for interaction and an optimal amount of feedback as constructive mediation in formative assessment supporting the learning process of inferential reading skills.

Readers see a glimpse of future possibilities for innovative assessments in Zourou's (2014) review, which attempted "to bridge game-based learning and game-based assessment, particularly in assessing complex problem-solving processes and outcomes in a digital game-based learning environment" (pp. 47–48). Zourou applied a computer-assisted language learning (CALL) perspective to the research presented in this book, which did not focus on language learning. The measuring progress, formative feedback, and social aspects of game-based learning are directly applicable to task-based language learning. The relationship between assessment and learning has implications for assessment of game-based learning in non-game and network-based learning contexts. Two useful concepts are the transfer of knowledge to real-world environments and the encouragement of foreign or second language interaction.

**Language Constructs in Technology-Mediated Environments**

Language teachers and testers working with new technologies in innovative ways inevitably encounter questions about how the abilities they teach and assess differ when language use is mediated through technology, and how to interpret the performance data gathered by the computer. García Laborda (2007) suggested in his review paper the need to "design new types of items for computers, especially for Internet-based tests" (p. 8). What kind of items should be created, on what basis should they be designed, and how should their success be evaluated? These are all questions that are typically addressed in large part on the basis of the construct that the assessment is intended to measure. *LLT* has published several papers that grapple with issues of construct definition because of the affordances that technology provides for designing assessment tasks and for analyzing test takers' language.

With respect to task design, *LLT* published a paper reporting research investigating the use of multimedia on a test of listening comprehension. Wagner (2007) investigated questions that have been raised by developers of listening tasks who are now working with computer-assisted test delivery, which makes the use of visuals including video an option for listening tasks. Such questions include the nature of the construct that is measured in a video listening test versus an audio-only listening test, the individual differences in the use of video materials across users, and the utility of the video in comprehending the meaning of the audio. By carefully examining the viewing the behavior of 36 students enrolled in a community English program, Wagner investigated the extent to which the students watched the video while they were completing listening comprehension tasks. He found that across task types, and throughout the test, learners tended to watch 69% of the time, suggesting that the videos were actually used. This descriptive research on test takers' behavior sheds light on the construct of listening as it operationalized during test taking. Test developers then need to consider the meaning of the test scores,

which should be treated as indicators of listening with visual support.

Another issue in task design was presented by Kol and Schcolnik (2008), who described research investigating the use asynchronous online discussion forum tasks as one form of assessment in a course for English for academic purposes. Participation in an asynchronous online discussion forum is a common activity found in many course management systems. Teachers can simply count instances of participation as a form of record keeping for students' participation, but Kol and Schcolnik (2008) showed how such record keeping might be transformed into a more meaningful and informative assessment through the development of evaluation criteria to assess the quality of the contributions rather than simply counting their presence and absence. The authors demonstrated that the process of developing meaningful criteria and scoring rubrics requires the teachers to specify what they are hoping to see in the students' contributions. In their study, this process resulted in the need to define the constructs of interaction, reflection, language complexity, and task purpose. Doing so increased teachers' understanding of their own goals in having students participate in online discussion.

With respect to response analysis, Crossley and McNamara (2013) used automated text analysis tools on transcribed spoken responses to tasks on a test of English for academic purposes. Their goal was to identify linguistic variables other than phonological ones that should be considered as part of the speaking construct measured by the test. The methodology used the ratings from the human raters as the dependent variable and the linguistic features of the spoken responses as independent variables. The linguistic features such as vocabulary size, causality, and word frequency predicted over half of the variance in test scores even without any phonological features such as phonological accuracy, intonation, and stress. Such findings help to provide evidence about the nature of the speaking construct as it is measured by human ratings and to provide baseline data for development of computer-assisted responses analysis as well.

**Impacts of Innovation**

One of the important areas of research on language assessment is the impact or consequences of language assessment on all stakeholders in the testing process and beyond. Accordingly, the review paper by García Laborda (2007) identified the study of the impacts of computer-assisted testing as one of the areas in need of future research. Many language testing researchers would argue that such research should be included in a program of validation required for justifying the interpretations and uses of test scores. García Laborda pointed out, however, that as of 2007, there were still very few washback studies investigating the "effect of computer-based tests on how teachers change their instruction style according to the computer interfaces in standardized tests, and the results obtained in this type of exams" (p. 8).

A second area of language test impact that language testers see as important is the effect of tests on learners' test preparation behaviors. The language learners' test preparation behavior is of interest because the test and the information provided to test takers should ideally encourage preparation behavior that improves the students' language skills. García Laborda (2007) imports this traditional concern from language testing to the reality of Internet communities where "test takers [can] communicate test strategies and information among themselves. These communities, if culturally based, could help many test takers overcome a lack of knowledge of the Anglo-European culture that underlies most of these tests" (p. 8). The study of test preparation on the Internet raises many important and interesting concerns for language testers—all of which come into play in language testing research.

A third type of impact that García Laborda (2007) suggested investigating is the extent to which test takers enjoy test taking. He introduced this as a new and unexplored issue, even though language testers have long been concerned about making test takers feel comfortable (e.g., Kenyon & Malabonga, 2001; Chen & Cheng, 2008) and placing them in a situation where they have the potential to demonstrate their best ability. The idea of enjoyment, however, may be an important dimension of an innovative technology agenda, where the use of feedback, individualization, and some ideas from online games might all play a role in creating enjoyable assessments for learners.

### Authoring Tools

An innovative agenda requires that many professionals be able to contribute to developing innovations. This need, in turn, creates a demand for authoring software that allows applied linguists to participate without requiring them to become computer programmers. Authoring tools are software applications designed to allow language teachers, materials developers, and language program administrators to create a variety of interactive computer-based assessments with embedded media, automated scoring, feedback, and database archival storage of performance (Kessler, 2013). Kessler pointed out that "authoring tools used in professional test development require technical expertise and have a steep learning curve" (p. 1). Tools for classroom use, however, often provide templates for non-programmers so that content can be easily added to traditional item types such as multiple-choice, matching, and fill-in-the-blank. Page layout and formatting are customizable in some programs while others require some coding to change appearance. In short, the term *authoring tool* encompasses a range of software designed for different purposes, but which can be used for developing language assessments.

Authoring tools for language assessment first gained attention in *LLT* in 2001 with a review article and two software reviews. The review article by Godwin-Jones (2001) presented fundamental technical issues of authoring tools at that time such as maintaining security, platform compatibility issues, standardizing questions, and test formatting. The author's prediction that authoring tools would become more flexible in their customization of specific feedback to meet the needs of individual classes or students was integral to increased usefulness. A decade later, Kessler (2013) praised customization options in more recent sophisticated authoring systems. Although the article by Godwin-Jones (2001) was a review of authoring tools and the technology needed for computer-assisted language assessment, most tools were initially developed to support language learning rather than to assess language ability.

In 2001, authoring systems dedicated specifically to the development of language assessments were scarce (Polio, 2001). The two reviews of authoring software in *LLT* presented practical issues regarding the adaptation of language learning software to the needs of language assessment. Benefits of adapting such software were presented in the review of Hot Potatoes software by Winke & MacGregor (2001). Hot Potatoes is a free online authoring suite with templates for six types of quizzes, which can be administered online, integrated into a content management system, or offered as a stand-alone program. No programming knowledge is required to add multimedia, edit scoring criteria, or customize feedback for correct or incorrect responses. The main challenge with this system, however, is test item security. Because of the likelihood of cheating, the authors advised against using this system for high-stakes testing.

An alternate commercial authoring tool, Test Pilot, was also reviewed by Polio (2001). Test Pilot has many of the same benefits as Hot Potatoes as well as item banking and the capacity to create computer-adaptive assessments. Both software tools provide customizable assessments and feedback in different ways, laying the foundation for Godwin-Jones' (2001) outlook on customizable individualized feedback. As an alternative to Hot Potatoes, the high cost was seen as the most prohibitive feature of this software for individual language instructors. These three reviews highlighted practical and technological benefits as well as features that need to be considered to improve authoring tools for language testing.

## CONCLUSION

This review should demonstrate how interwoven technology and language assessment are with language learning. Many of the same basic technologies that play a role in language teaching are also put to work in language testing; but in language testing, they can take on new meaning. Grasping their meaning requires knowledge of basic concepts in language assessment, which to date have not become common currency among language teachers and other professionals in applied linguistics. Professionals in language assessment are keen to expand knowledge of testing concepts to allow more engagement in language

assessment across the profession, and technology exacerbates this need. In the "On the net" column in *LLT*, LeLoup and Ponterio (2001) described an online resource intended for professional development. The website, maintained by Glenn Fulcher, contains a composite of video clips on topics in language testing, reviews of websites related to language testing, a bibliography of language testing articles, working papers, research reports, and a database of current research projects. This website has been redesigned and updated since the 2001 review and continues to offer relevant information for the language testing community.

Professional development and teacher education are key areas of interest for those wanting to develop best pedagogical practices on how to integrate useful assessments in a CALL environment. As the capability of computers evolves and language learners enter the classroom with computer skills, teachers need to reconsider how computers can be used effectively in a second language classroom (Chapelle & Jamieson, 2008). Articles in *LLT* have demonstrated some of the benefits and challenges of using technology for language learning and assessment. For example, the range of attitudes and performance results varied based on the different ways of integrating MyAccess! into a language course (Chen & Cheng, 2008). It was noted that, "writing teachers need to be fully aware of the limitations of AWE technology as well as students' learning needs and contexts in making decisions about how to maximize effective AWE use and to minimize undesirable outcomes" (p. 110). Technology training courses in teacher education programs, however, are frequently taken as electives outside the department or too late in the program to benefit the student (Hegelheimer, 2006). We hope to see increased attention in teacher education programs on the role that technology plays not only in language learning but also in second language assessment.

Another area where there is room for progress is in speech recognition. We have seen that measuring oral proficiency using computer technology is an area of research interest in *LLT* (Blake, Wilson, Cetto, & Pardo-Ballester, 2008; Kenyon & Malabonga, 2001). The ability to automatically score oral performance begins with recognition of speech. Yet, due to the developmental nature of learner speech, recognition engines are not yet sophisticated enough to accommodate all levels of learner language. Automatic speech recognition has thus far been limited to short answers and constrained responses. Some degree of success has been achieved in generating scores for longer speech samples (Zechner, Higgins, Xi, & Williamson, 2009), and we look forward to incorporating more accurate and reliable speech recognition in second language learning and assessment in the future.

However, as Brown (1997) predicted in his review of technology and language assessment in the first issue of *LLT*, the technology-related issues of language assessment will continue to increase in their detail and complexity, and the need will continue to grow for professionals capable of negotiating the many considerations that come into play in the design of computer-assisted language tests. With so little written about technology use in language assessment, much more is needed at this intersection. It is an area where *LLT* has filled some of the gap in the past, and undoubtedly will continue to contribute in the future.

## ABOUT THE AUTHORS

Carol Chapelle is a Distinguished Professor in the Program of Applied Linguistics and Technology at Iowa State University where she teaches courses in second language acquisition, language testing, and computers in applied linguistics. She has published extensively in the areas of computer-assisted language learning and language assessment.

**E-mail**: carolc@iastate.edu

Erik Voss is an Associate Teaching Professor at Northeastern University. His research focuses on validation research, corpus linguistics, and language assessment and technology. He has presented on technology use in the language classroom and language assessment research at domestic and international conferences.

**E-mail**: e.voss@neu.edu

## REFERENCES

Blake, R., Wilson, N. L., Cetto, M., & Pardo-Ballester, C. (2008). Measuring oral proficiency in distance, face-to-face, and blended classrooms. *Language Learning & Technology*, *12*(3), 114–127. Retrieved from http://llt.msu.edu/vol12num3/blakeetal.pdf

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology*, *1*(1), 44–59. Retrieved from http://llt.msu.edu/vol1num1/brown/

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

Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language Learning & Technology*, *5*(2), 95–98. Retrieved from http://llt.msu.edu/vol5num2/deville/

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

Chapelle, C. A., & Jamieson, J. (2008). *Tips for teaching with CALL: Practical approaches to computer-assisted language learning.* New York, NY: Pearson Education.

Chen, C.-F. E., & Cheng, W.-Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12*(2), 94–112. Retrieved from http://llt.msu.edu/vol12num2/chencheng.pdf

Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*(2), 171–192. Retrieved from http://llt.msu.edu/issues/june2013/crossleymcnamara.pdf

Dunkel, P. A. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology*, *2*(2), 77–93. http://llt.msu.edu/vol2num2/article4/

Fernández-García, M. (2001). [Review of the book *Issues in computer-adaptive testing of reading proficiency* by M. Chalhoub-Deville (Ed.)]. *Language Learning & Technology*, *5*(2), 19–22. Retrieved from http://llt.msu.edu/vol5num2/review1/

García Laborda, J. (2007). On the net: Introducing standardized EFL/ESL exams. *Language Learning & Technology*, *11*(2), 3–9. Retrieved from http://llt.msu.edu/vol11num2/net/

Godwin-Jones, R. (2001). Language testing tools and technologies. *Language Learning & Technology*, *5*(2), 8–12. Retrieved from http://llt.msu.edu/vol5num2/emerging/

Godwin-Jones, R. (2008). Web-writing 2.0: Enabling, documenting, and assessing writing online. *Language Learning & Technology*, *12*(2), 7–13. Retrieved from http://llt.msu.edu/vol12num2/emerging.pdf

Hegelheimer, V. (2006). When the technology course is required. In P. Hubbard & M. Levy (Eds.), *Teacher education in CALL* (pp. 117–33). Amsterdam, Netherlands: John Benjamins.

Kessler, G. (2013). Authoring tools for language assessment. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Oxford, UK: Wiley-Blackwell.

Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology, 5*(2), 60–83. Retrieved from http://llt.msu.edu/vol5num2/kenyon/

Kol, S, & Schcolnik, M. (2008). Asynchronous forums in EAP: Assessment issues. *Language Learning & Technology, 12*(2), 49–70. Retrieved from http://llt.msu.edu/vol12num2/kolschcolnik.pdf

LeLoup, J. W., & Ponterio, R. (2001). On the net: Language testing resources. *Language Learning & Technology, 5*(2), 4–7. Retrieved from http://llt.msu.edu/vol5num2/onthenet/default.html

Lim, H., & Kahng, J. (2012). [Review of the software CRITERION]. *Language Learning & Technology, 16*(2), 38–45. Retrieved from http://llt.msu.edu/issues/june2012/review4.pdf

Mislevy, R., Chapelle, C. A., Chung, Y.-R., & Xu, J. (2008). Options for adaptivity in computer-assisted language learning and assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 9–24). Ames, IA: Iowa State University.

Norris, J. M. (2001a). [Review of the book *Computerized adaptive testing: A primer* (2nd ed.) by H. Wainer (Ed.)]. *Language Learning & Technology, 5*(2), 23–27. Retrieved from http://llt.msu.edu/vol5num2/review2/

Norris, J. M. (2001b). Concerns with computerized adaptive oral proficiency assessment [Peer commentary on the article "Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments" by D. Kenyon & V. Malabonga]. *Language Learning & Technology, 5*(2), 99–105. Retrieved from http://llt.msu.edu/vol5num2/pdf/norris.pdf

Polio, C. (2001). [Review of the software TEST PILOT]. *Language Learning & Technology, 5*(2), 34–37. Retrieved from http://llt.msu.edu/vol5num2/review4/

Roever, C. (2001). Web-based language testing. *Language Learning & Technology, 5*(2), 84–94. Retrieved from http://llt.msu.edu/vol5num2/roever/

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology, 5*(2), pp. 38–59. Retrieved from http://llt.msu.edu/vol5num2/sawaki/

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge.

Teo, A. (2012). Promoting EFL students' inferential reading skills through computerized dynamic assessment. *Language Learning & Technology, 16*(3), 10–20. Retrieved from http://llt.msu.edu/issues/october2012/action.pdf

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology, 11*(1), 67–86. Retrieved from http://llt.msu.edu/vol11num1/wagner/

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Winke, P., & MacGregor, D. (2001). [Review of the software Hot Potatoes]. *Language Learning & Technology, 5*(2), 28–33. Retrieved from http://llt.msu.edu/vol5num2/review3/default.html

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning & Technology*, *8*(1), 53–65. Retrieved from http://llt.msu.edu/vol8num1/wolfe/

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883–895.

Zhang, L. (2014). [Review of *Handbook of Automated Essay Evaluation: Current Applications and New Directions* by M. D. Shermis & J. Burstein (Eds.)]. *Language Learning & Technology*, *18*(2), 65–69. Retrieved from http://llt.msu.edu/issues/june2014/review2.pdf

Zourou, K. (2014). [Review of *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives* by D. Ifenthaler, D. Eseryel, & X. Ge (Eds.)]. *Language Learning & Technology*, *18*(3), 47–51. Retrieved from http://llt.msu.edu/issues/october2014/review3.pdf

Language Learning & Technology. A refereed journal for second and foreign language scholars and educators. Home.Â  The inaugural Dorothy Chun Award for Best Journal Article in Language Learning & Technology has been granted to Dr. Ines Martin. Read more about her achievement and this new format for recognizing research excellence in LLT. COVID-19 Delays. 20 years of technology and language assessment in. Language Learning & Technology, 20 (2), 116â€“128. Google Scholar. Clariana, R., & Wallace, P. (2002).Â Supervised learning of universal sentence representations from natural language inference data. Volume~1. arXiv preprint arXiv:1705.02364. Croft, A. C., Danson, M., Dawson, B. R., & Ward, J. P. (2001). Experiences of using computer assisted assessment in engineering mathematics. Computers & Education, 37 (1), 53â€“66. Google Scholar. Dumais, S. T. (2005). Latent semantic analysis. Annual Review of Information Science and Technology, 38 , 188â€“230. Google Scholar. Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., & Dang, H. T. (2013). Assessment of learnersâ€™ language ability is an important part of language education, which has been affected by computer technology at least as significantly as language learning has. Because of the significance of language assessment for language teachers, software developers, applied linguists, and learners, articles in Language Learning & Technology (LLT) have contributed to chronicling the developments in language assessment technologies. Throughout these papers, the terms language testing and language assessment are used to denote the process of systematically gathering data from learners...