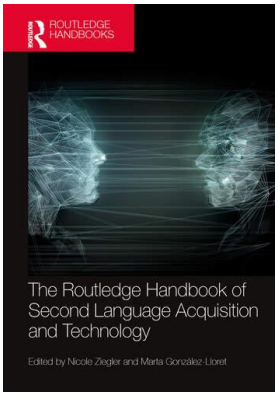

The Routledge Handbook of Second Language Acquisition and Technology

Nicole Ziegler, Marta González-Lloret

Technology and Assessment

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781351117586-10
Phuong Nguyen, Volker Hegelheimer
Published online on: 01 Feb 2022

8

TECHNOLOGY AND ASSESSMENT

Phuong Nguyen & Volker Hegelheimer

Introduction

Since its early inclusion in language teaching and assessment in the 1960s, technology has rapidly developed to become more versatile and capable than ever before. This advancement has led to assessment innovations that include enhanced test development and delivery, increased accessibility, and reduced test score reporting time. Part of the motivation for technology integration, particularly in large-scale language testing, stemmed from the need to make the testing process more efficient. This integration also challenged language testing experts to refine definitions of test constructs. Furthermore, technology-assisted assessment has led to some level of reliance on automated scoring, which was pioneered in the assessment of writing and has now become more prevalent in speaking assessments as well.

In this chapter, we will first provide a brief history of technology use in language assessment. Then, we will present several issues and topics that are critical in technology-mediated language assessment, including construct, test security, and washback considerations. We then outline contributions of technological innovations to second language (L2) assessment as well as current research and major research methods in this area. We further provide recommendations for language test developers and instructors in terms of practice. We conclude the chapter with notes on future directions for the use of new technologies in L2 assessment.

Historical Perspectives

The introduction of technology in testing can be traced back to 1935, when an IBM computer was first used in the U.S. to score multiple-choice tests. With rapid technological changes during the 1960s, language testers, motivated by the need to make the testing process more efficient, began to use technology for many purposes, including analyzing test data, storing test items in databases, and producing test result reports for test users. Efforts to understand L2 computer-based testing (CBT) emerged in the mid-1980s with the 1985 Language Testing Research Colloquium (LTRC) and the publication of the conference proceedings by Stansfield (1986). The 1990s saw the emergence of L2 computer-adaptive testing (CAT), which applies latent trait models to select and present test items in a sequence based on the examinee's responses to each item. In fact, many universities and major language testing agencies were engaged in developing CBT or CAT systems at this time. Examples of such systems include the Montgomery County Public Schools project (Stevenson & Gross, 1991), Brigham Young University CAT instruments for placement tests of French, German,

DOI: 10.4324/9781351117586-10

and Spanish (Madsen, 1991), CommuniCAT and CBT BULATS developed by the University of Cambridge Local Examinations Syndicate (UCLES), and the CBT Test of English as a Foreign Language (TOEFL) released by the Educational Testing Service (ETS). Also, in this period, the development of the Internet allowed for web-based testing (WBT), in which the test, located as a website on the tester's server, can be accessed by the examinee's computer.

Currently, many language tests are delivered online. For example, the DIALANG project, which is sponsored by the Council of Europe to provide diagnostic assessment in 14 languages, as well as the TOEFL iBT and CBT BULATS, are all available as online assessments. Various technologies are used for test administration and delivery. For instance, web-based applications (e.g., Extempore, Learnosity, and Questionmark Perception) have been developed as assessment management systems specializing in assessment solutions for educational purposes. These platforms allow for question authoring, item bank management, and test design, and additionally serve as educational, multimedia platforms. Also, video-conferencing tools (e.g., Zoom and Adobe Connect) have been employed for L2 oral proficiency. Dialogue-based systems, in their many forms such as intelligent tutoring systems, games and virtual environments, spoken dialogue systems, and chatbots, have been used for language assessment, especially in the classroom context or language for specific purposes (Park, 2018; Ramanarayanan et al., 2018). Additionally, language test delivery is no longer limited to the computer but has also expanded to mobile devices, as in the case of Pearson's tablet-based English Benchmark Young Learners or Duolingo's adaptive speaking test, which can be taken on examinees' cell phones or tablets.

Advancements in natural language processing have allowed for automated scoring of learners' performance and feedback to their responses in high-stakes and low-stakes testing. Informed by educational measurement, computer science, and linguistics, such systems employ both statistical and linguistic methods to analyze relevant linguistic evidence in the examinees' responses and then generate a mapping function from a combination of linguistic evidence to a score similar to that obtained from human raters. Automated writing evaluation (AWE), introduced in the 1960s to promote language teaching and assessment efficacy, has been used for evaluating examinees' performance by many language-testing organizations. Three examples of such systems include e-rater®, Intelligent Essay Assessor®, and IntelliMetric™, developed by ETS, Pearson Knowledge, and Vantage Learning, respectively, to assess written responses to their tests. More recently, automated speech recognition (ASR) systems have been developed and utilized to score spoken language. To date, the most advanced ASR systems that demonstrate the current state-of-the art in language assessment are SpeechRater™, owned by ETS and designed to score English speech, and the Versant testing system, owned by Pearson and used to score speech of languages other than English, such as Arabic, Dutch, French, and Spanish.

The examples above illustrate how L2 assessment has progressed to integrate more technological advances in many aspects of the language test development cycle. In addition to enhancing efficiency such as automation of existing practice in L2 assessment, these technologies have important implications for the interpretation and use of test scores and for L2 learning, all of which are central issues in language assessment research.

Critical Issues and Topics

As technology is increasingly integrated into many aspects of language assessment, language assessment researchers have discussed opportunities and challenges presented by these technologies. Some critical issues and topics that interest language assessment researchers pertain to the construct of language ability, test security, and washback effects.

The central issue that has attracted attention from language assessment experts is the issue of how to define constructs. Some challenges in terms of construct definition have arisen since the advent of computer-based language testing. First, the implementation of multimedia in language

tests results in a more complex construct to measure, which, in turn, poses a threat to test validity (Douglas & Hegelheimer, 2007). For instance, performance on a computer-based writing test might be affected by examinees' typing proficiency. Similarly, as the growth of multimedia has allowed for the use of visuals in listening tests, researchers have questioned whether test constructs should be limited to aural listening ability. In technology-mediated language tests, test developers and researchers have also tried to identify potential construct-irrelevant variables inherent to the use of technology. The second challenge pertains to the use of automated scoring systems as they affect the construct being measured by the test. Although to some extent automated scoring systems can help eliminate the inherent subjectivity in human scores, which may increase the reliability and consistency of the scoring process, the limitations of current AWE and ASR systems narrow the construct being assessed (i.e., construct underrepresentation) and thus, threaten the validity of the claims made based on the test results. For example, in ASR systems, it is challenging to capture the complexities of speaking performances because the current systems, mainly developed to score monologic speech, are not efficient in capturing aspects such as content, appropriateness of word choice or expression use, and development of ideas. In such a case, the argument that test-takers can communicate appropriately in the target language cannot be substantiated when such an ASR system is used as the sole rater for the test. A similar caution applies to the Duolingo English Test, a computer-adapted speaking test administered on the computer or mobile devices using the device's screen, camera, keyboard, speakers, and microphone. The test, which claims to assess general L2 speaking ability, requires examinees to read aloud a written sentence presented on the screen. As the test is computer-adaptive, examinees' speech sample is scored automatically and instantly by the computer using a proprietary algorithm. However, language assessment experts (e.g., Wagner, 2020) have cautioned that this test measures a rather limited construct of English spoken proficiency because of the nature of the speaking task (i.e., sentence read-aloud).

Additionally, the fact that tests can be delivered on the computer or mobile devices on the Internet engender concerns about test security. Technological integration, especially in high-stakes testing, might facilitate cheating or hacking of the testing systems. The growth and increasing sophistication of high-tech gadgets, such as digital recorders, smart watches, screen capture applications, might ease the process of cheating or recording test item information. The use of at-home tests, such as the Duolingo English Test, has raised questions regarding cheating and thus, test score reliability. It should be noted that some language testing agencies, such as ETS, have developed sophisticated software and algorithms to analyze examinees' response patterns for collusion. However, so far, very little research has been conducted to investigate the issue of cheating in at-home Internet-based tests. This topic is likely to continue to be an area of interest to language assessment developers and researchers.

One important area of research in language assessment is its impact on all stakeholders, including students, teachers, and educational administrators. When a speaking test requires examinees to repeat a series of sentences and produce short sentences, which are then scored automatically, what implications does it have for language teaching and learning? Or if a test is delivered on mobile phones, does it change the way language teachers teach and how students learn? With the increased use of technology in L2 assessment, it is undoubtedly essential to investigate its effects on L2 teaching and learning (also called test washback) because any innovation in language assessment should give examinees opportunities to learn from both the process and the results of test taking (Chapelle & Voss, 2014, p. 121). Therefore, one topic of interest to language assessment researchers, although little research has been conducted to date, is how language teachers change their teaching practices in response to computer use in assessment, especially standardized assessment. It is also important to study the effects of tests on learners' test preparation, which could provide useful validity information about tests because technology-mediated language tests, like other types of assessments, should encourage test preparation behavior that improves learners' language skills.

Current Contributions and Research

A plethora of research articles in language testing have also focused on technology use in L2 assessment. Overall, contributions of technology in L2 assessments are complex and varied. However, they can generally be categorized in terms of practicality, reliability, authenticity, construct, and washback effect.

Practicality

Technological integration has made it more convenient and less labor-intensive to administer L2 assessments and to score examinees' performance, which consequently offers financial benefits to language testing organizations. For instance, the use of video-conference tools or virtual environments helps overcome the issue of examinees and examiners who are not conveniently available in the same location and reduces the cost for testing venues. Although labor-intensive and expensive to develop, AWE and ASR systems for scoring written and spoken responses make scoring more time efficient and in the long run, become more cost effective because of potential reduction in the cost of hiring and training human raters. These benefits help reduce the cost not only for testing organizations but also for examinees. Additionally, since the outbreak of the COVID-19 pandemic, the importance of technology for L2 assessment has been demonstrated, both in large-scale testing and in classroom contexts. For example, IELTS Indicator was launched in late April 2020 and has been used as an online substitute for the regular IELTS Academic.

In the language classroom context, technology can also potentially enhance classroom assessment. For instance, AWE technologies, such as Criterion (ETS's web-based service that uses e-rater® to assess students' writing and provides instantaneous score reporting and diagnostic feedback) could be used to assess grammar and mechanics. Thus, teachers' attention would be directed more to content and organization, which allows them to assess students' work more efficiently and with greater care. In addition, students would receive similar or greater volume of feedback overall. Other technological efficiencies, like VoiceThread and Zoom, might make good, structured, and regular classroom assessment more practical. In fact, VoiceThread can be efficient for speaking formative assessment because of its ability to allow learners to interact with each other using image, voice, and videos at any time. There is no doubt that technology has made test delivery and scoring more efficient and practical through its availability and ability to provide automatic scoring and feedback.

Reliability

Another benefit of technology for language assessment pertains to test score reliability, which can be improved by the use of automated scoring of examinees' responses and consistent test administration. The use of AWE and ASR systems for scoring examinees' responses helps enhance rating consistency, which is hard to achieve with human raters as raters can be a source of test score variability (Barkaoui, 2010). In fact, research on well-known AWE systems has shown that these systems can be more consistent compared to human raters. For example, in the case of Intelligent Essay Assessor® (IEA), an AWE system developed by Pearson Knowledge to provide summative assessment in combination with tutorial feedback to examinees' writing, the correlations between scores provided by IEA and those given by human raters (i.e., IEA to human reliabilities) are very similar to or even higher than the correlations between scores provided by two human raters (i.e., human to human reliabilities) (Foltz et al., 2013). A similar conclusion has also been reached for e-rater®, a system developed by ETS and used as a second rater to complement human scoring for many writing tests including the TOEFL iBT (Attali & Burnstein, 2005). Thus, automated scoring can help maintain score consistency or reliability across items and over time. Furthermore,

technology allows for consistent test administration because it minimizes the effects of such variables as training of human administrators and environmental factors. This will undoubtedly contribute to the validity of test score interpretation and use and enhancing test fairness among examinees.

Authenticity

However, practicality and reliability are not the only areas that benefit from technology use in L2 assessment. In fact, technology also enables the development of innovative, authentic task types. For instance, the integration of videos in listening tests enhances test authenticity. Listening with visuals, where the listener can see non-verbal features from the speakers, is more similar to real-life listening tasks than listening tests that rely on aural only input (Ockey, 2007; Wagner, 2010). When employed for classroom-based speaking assessment, dialogue-based systems, in their many forms such as intelligent tutoring systems (see Chapter 9 this volume), games and virtual environments (see Chapters 20, 21, 23 in this volume), spoken dialogue systems, and chatbots, can promote design of authentic tasks. Thanks to new technologies that can provide examinees with stimuli as audio or video files, test developers can include integrated speaking tasks similar to real-world communication tasks in which speaking often involves responding to some input or understanding visual cues from other speakers. This provides a more robust, authentic medium for L2 speaking assessment. Similarly, interactive virtual tasks, such as those developed for aviation English tests to replicate the real-world scenarios that air traffic controllers encounter in their daily job (Park, 2018), could be useful for assessing English for Specific Purposes in other fields, such as business and tourism. The fact that new technologies can facilitate the design of authentic tasks is undeniably a significant benefit for L2 assessment, as examinees' language proficiency should be ideally assessed in an authentic task and situation (Douglas, 2013).

Construct

Another important contribution of technological innovations, which is closely related to authenticity of examinees, is related to the construct of language ability. First, innovative, authentic test tasks, developed with the help of technology, allow test developers to measure aspects of the language ability construct that were previously difficult or impossible to assess in previous decades. For example, technology can foster the design of innovative speaking tasks that can measure unexplored L2 territory such as interactional competence in computer-based testing. With the use of video-conferencing tools, language testers can measure spoken proficiency of many examinees concurrently and create tasks that require discussion and collaboration among these examinees, allowing examiners and examinees to access the vocal and facial cues of their interlocutor. This allows for a more authentic assessment of interactional competence, which includes examinees' ability to comprehend spoken input and appropriately produce language in response to the input by negotiating meaning and taking turns, an important aspect of oral proficiency (Louma, 2004; Ockey, 2018). Although interactional competence can be assessed using face-to-face speaking tasks, it can be assessed more conveniently with the help of video-conferencing technologies or dialogue-based systems. The capabilities of automated scoring systems further illustrate this point that technology can assist with a more comprehensive assessment of the construct. These systems, such as IntelliMetric™, owned by Vantage Learning, or ETS's e-rater® and SpeechRater™, can analyze hundreds of linguistic features simultaneously (Attali & Burnstein, 2005; Chen et al., 2018; Schultz, 2013). This would be an otherwise impossible task for human raters who might suffer from cognitive overload from rating with technology because of fatigue (Knoch et al., 2007) or multitasking (i.e., navigating the testing interface while rating, and resolving technical issues that might arise during test administration). Therefore, given that second language acquisition (SLA)

helps identify language components for elicitation and criteria assessment (Shohamy, 2000), technology allows the construct of language ability to be more fully operationalized in tests.

Second, technological innovations have contributed to the understanding of the construct of language ability. Some studies have found that the use of technology in language tests enables the construct of language ability to be re-defined. For example, the use of multimedia in listening assessment requires a broader construct of listening ability, which includes not only aural listening ability but also the ability to use visual cues (e.g., facial expressions and hand gestures) to understand the aural text (Ockey, 2007; Wagner, 2008, 2010). Thus, findings in technology-mediated language assessment could potentially be used to develop models of language ability, which according to Shohamy (2000) can be useful for SLA researchers in examining the validity of their findings.

Washback

More importantly, technology-mediated assessments can encourage learning because of their unique capabilities. First, they are able to offer individualized analysis of learners' language, feedback, and reporting, which consequently encourages learners to make decisions about the feedback and evaluate their learning, that is, assessment *as* learning (see Dann, 2014). This benefit is exemplified by the use of AWE systems such as Criterion, the Writing Mentor™, and MY Access!®, a web-based writing assessment tool relying on IntelliMetric™. In a study on Criterion, Attali (2004) found that students were able to understand and attend to the feedback provided by the system and more importantly, to improve the quality of the essay beyond the mere lengthening of it. Heffernan and Otoshi (2015), finding that students receiving feedback from both a teacher and Criterion showed improvement in some rhetorical features, concluded that Criterion, along with human guidance and feedback on early drafts, can be useful for EFL learners' writing development. Evaluation studies on the effectiveness of the Writing Mentor™, a Google Docs add-on developed by ETS targeting struggling writers and college-level English learner populations, have shown that it has generally received positive feedback from users and that the tool has led to writers' revision of their draft (Burnstein et al., 2018). MY Access!® is another AWE system that aims to provide students with a writing environment where they receive both immediate scoring and formative feedback so that they can improve their writing proficiency accordingly. This system provides multi-lingual support and various types of feedback, such as focus, organization, content development, language use, and mechanics and conventions, along with other support such as multilingual dictionaries, translators, and thesauruses (Vantage Learning, n.d.). Also, instructors can utilize various functions embedded in the system to add their feedback to students' writing, organize groups and customize the type of feedback according to students' proficiency level, and generate reports on their students' progress (Vantage Learning, n.d.). Examining how students and instructors adapted MY Access!® to a pedagogical tool that provides immediate computer-generated scores along with diagnostic feedback in three EFL college writing classes, Chen and Cheng (2008) concluded that the diagnostic feedback function of the tool seemed to be pedagogically appealing for formative learning. This combination of scores and diagnostic feedback from automated scoring systems, when used in combination with feedback from instructors with a sound pedagogical foundation, promises to support learning-oriented assessment, that is, the assessment *of* and *for* learning (Turner & Purpura, 2016). Those more pedagogically oriented tools have the potential of facilitating personalized learning and helping improve academic writing skills, and thus, could be used as part of writing instruction.

Second, technological innovation can help develop assessments that create opportunities for student learning beyond the bounds of possibilities in the traditional language classroom. For example, computer-assisted dynamic assessment could replace the role of a human mediator in human-to-human mediation for dynamic assessment. An action research project by Teo (2012) explained how a web-based computerized dynamic assessment program could provide opportunities for interaction

and feedback as constructive mediation in formative assessment to improve students' metacognitive reading strategies in making inferences. Yang and Qian (2020) also found beneficial effects of computerized dynamic assessment in promoting EFL learners' reading comprehension.

Finally, new technologies help construct assessments that could encourage learning by providing students with authentic and immersive learning environments and thus, potentially facilitate their motivation to learn the target language. When used for classroom assessment, games and virtual environments, which are increasingly sophisticated, interactive, and visually appealing, allowing for multiple interlocutors to communicate, cooperate, and compete through virtual spaces and avatars (Lin & Lan, 2015), make assessment more enjoyable for students and might motivate students to learn. In fact, Forsyth et al. (2019) found that their multimodal dialogue system developed for speaking assessment in academic scenarios generally received positive feedback from examinees. In addition to creating possibilities to assess examinees' proficiency in complex situations in the classroom, these systems, though expensive to develop or customize for educational purposes, also provide learning opportunities to students thanks to their ability to measure students' progress and give formative feedback, as well as the interaction students engage in these virtual realities. Therefore, technological innovations, if used appropriately and responsibly, can expand the resources and improve the efficiency of language learning and assessment.

Overall, technological innovations have made great contributions to L2 assessment, instruction, and to some extent, SLA research. In fact, technologies that help with automation of existing practice in L2 assessment, such as CAT, video-conferencing applications, and automated scoring systems, have made assessment more practical, efficient, and reliable. With their abilities to create authentic tasks, new technologies now allow for assessment of more aspects of the construct of language ability and provide opportunities for researchers to re-define the construct. Consequently, this potential for innovation will benefit SLA research as results about the construct of language ability can help SLA researchers to validate their findings regarding learners' performance and development. Finally, technology-mediated assessments can promote learning with their unique capabilities of giving instant feedback, as well as creating authentic and immersive learning environments that can foster students' motivation for language learning.

Main Research Methods

Research in the technology use in L2 assessment varies in their methods. Many validation studies of technology-mediated language tests adopt nonexperimental, *ex post facto* research designs in which the investigation starts after the event has occurred without interference from the researcher (Salkind, 2010). In these studies, data such as test scores or examinees' responses are normally provided to researchers who have no involvement with the testing process. Some examples of such studies include those that investigate examinees' linguistic features (e.g., Biber & Gray, 2013; Kyle et al., 2016), reliability of test scores provided by automated scoring systems (e.g., Attali & Burstein, 2005; Foltz et al., 2013; Rudner et al., 2006; Schultz, 2013), or concurrent validity of tests (Bernstein et al., 2010).

Experimental/quasi-experimental research methods are also utilized to identify the effect of a certain type of technology on students' performance. For example, experimental or quasi-experimental designs have been used to explore a variety of topics in L2 assessment, such as the effects of visual input on listening performance (Wagner, 2008), the impact of keyboarding skills on computer-based writing assessment (Barkaoui, 2014), and the influence of mode of assessment delivery on learners' performance on writing tests (Brunfaut et al., 2018). In the classroom context, quasi-experimental design has been used to study the effects of a computer-based dynamic assessment system on students' reading strategies and comprehension (Teo, 2012; Yang & Qian, 2020). Overall, experimental/quasi-experimental research designs allow researchers to control as many factors as possible to identify the effects of a certain technology on examinees' performance.

Although these studies mentioned above are mainly quantitative in nature, drawing on various statistical procedures for data analyses, recent research on technology-mediated language assessment has seen an increasing use of qualitative methods to investigate examinees' cognitive processes during technology-mediated L2 tests or their perceptions about such tests. For example, examinees' verbal reports were analyzed to investigate their attendance to non-verbal information in video listening tests (Ockey, 2007; Wagner, 2008). More recently, Tarighat and Khodabakhsh (2016) examined learners' attitudes towards Mobile-Assisted Language Assessment (MALA), by analyzing interviews from 17 advanced learners of English who reported on their views on the use of WhatsApp for assessing their speaking proficiency. Their findings demonstrated learners' mixed attitudes towards MALA because of concerns regarding fairness and authenticity. These studies have demonstrated that the qualitative research method can provide deeper insights into examinees' behaviors during technology-mediated language tests as well as their perceptions about such tests.

Many studies have recently adopted a mixed-methods (MM) design where both test scores and survey responses, as well as other types of more qualitative data, such as interviews, think-aloud protocols, and stimulated recalls, are collected to examine the effects of technology on test performance, examinees' perceptions and cognition, and rating behaviors. An example of MM research is a study by Kiddle and Kormos (2011), who examined students' spoken performance and perception in a direct face-to-face speaking test and in its online version on Questionmark. Analysis of their participants' test scores showed no significant difference in the difficulty of the two versions of the test while examinees' comments revealed their preference for the face-to-face version. MM research has also been used to examine the use of Adobe Connect for speaking assessment (Kim & Craig, 2012) and the use of Zoom for the International English Language Testing System (IELTS) speaking test (Nakatsuhara et al., 2017). Analyzing examinees' scores and linguistic output as well as examiners' verbal reports, Nakatsuhara et al. (2017) found that the Zoom test version generated similar test scores to the face-to-face version, despite of some differences in examinees' functional output and examiners' behaviors. Thus, the use of MM research approaches can enrich results from quantitative analyses, helping researchers not only identify the effects of the implemented technology on examinees' performance but also describe the complexity of the interaction between technology use and these stakeholders.

Overall, researchers may draw on a range of research methods when investigating various issues relating to technology use in language assessment. However, in line with the move to adopt MM in language assessment (Turner, 2013), researchers in technology-mediated language assessment have also taken advantage of the strengths of this method to provide more insights into the effects of technology on stakeholders and the validity of test scores interpretation and use.

Recommendations for Practice

The previous sections have identified some critical issues in technology-mediated L2 assessment as well as presented some of the major contributions of technology and research in the field of language assessment. In this section, we address the implications for practice for language testers and teachers.

The first implication is related to test construct. Since technology might be a variable influencing performance, when it is integrated in test task design and test delivery, test developers need to attend to the impact of the integration when defining the construct for their test. For instance, in a computer-based writing test, does computer literacy impact performance, and thus, should it be included in construct definition? In language tests where technology is involved, test developers should be clear about the construct they are measuring to help stakeholders make valid inferences about examinees' language ability.

Additionally, because of the possibility of dishonesty in technology-mediated tests, it is important that test developers have stringent protocols to maintain the security of their test materials, test delivery systems, and scoring systems, as well as to ensure that assessments are administered in secure, standardized environments. The test administration process should be monitored carefully by collecting test-takers' identity information onsite using photos, facial recognition, biometric scanning, etc. Webcams could be used to confirm the identity of registered examinees if the test is administered in locations other than the testing site. For language testing organizations with resources, AI algorithms could be employed to detect potential indicators of malicious behavior by attending to examinees' responses and their response patterns. Also, third-party proctoring services, which employ webcams to ensure appropriate home testing environment (e.g., ProctorU and Proctorio), could be an option for testing organizations and educational institutions. Plagiarism checkers might also be useful for both high-stakes and classroom/learning-oriented assessment. These procedures help ensure reliable test scores that provide valid information about examinees.

In the language classroom context, technology can be utilized for both formative and summative assessment purposes. For instance, video-conferencing technologies, which have become more popular for remote instruction, could be used to assess students' oral communication skills during group discussions or as a final exam. Automated scoring systems could be used in conjunction with teachers' instruction for class assignments so that students can use the feedback provided to improve their language skills. Teachers should also encourage repeated practice using automated technology-delivered assessments so that learners can engage in ample practice and receive automated feedback. This would also allow teachers to focus their feedback on aspects of learner production that technology does not capture well. Additionally, teachers could make use of their institution's learning management system to create test items and deliver tests to students for formative, low-stakes assessments. In cases where the assessment is high-stakes, rigorous methods of maintaining security should be adopted to help ensure that test results reflect students' language ability.

Future Directions

In the future, new technologies will continue to evolve and transform language assessment. First, performance-based language tests are likely to be entirely technology-mediated as tasks can be simulated easily with the help of technologies such as video-conferencing technology and dialogue systems. Test tasks will become more authentic because of the development of increasingly sophisticated and mature technologies, such as authoring tools and speech recognition technology, which allow for the integration of multimedia stimuli and virtual environments in test design and administration. Second, automated scoring systems will continue to be perfected and employed to score examinees' responses in high-stakes testing and to provide formative feedback in classroom contexts. They will also be able to extract linguistic features that are not effectively captured by current systems, such as idea organization or lexical appropriateness. ASR systems will be more accurate and reliable when analyzing examinees' acoustic input, especially that of nonnative accents, as well as open-ended, longer responses. Additionally, with the development of machine learning and natural language processing, it is likely that future generations of assessment will be able to make predictions about students' progress and calculate trajectories in language learning, thus providing useful information to language educators who can then personalize their instructions to individual students.

In terms of research, because of the potential advantages and challenges presented by emerging technologies in L2 language assessment, more studies are needed to explore the effects of technological integration on the validity of test score interpretation and use. First, more research should be conducted to understand the constructs measured by technology-mediated L2 tests to

examine whether test constructs are under-represented or if construct-irrelevant variables exist. As Brunfaut et al. (2018) recently identified differences in cognitive load among examinees depending on the delivery mode of the assessment (paper-based vs. online), the issue of construct-irrelevant variables should be addressed for every technology-mediated assessment. Studies on test constructs in technology-mediated L2 assessment will continue to shed light on the language ability models, which will benefit SLA researchers in validating their findings (Shohamy, 2000). Second, researchers should also investigate the effects of such assessments, especially with the use of automated scoring systems, on language learning and instruction to understand the use of assessment for increasing and improving opportunities for learning. For example, how does the integration of technology-mediated assessment change teachers' instruction style and students' learning behavior, such as their learning autonomy and motivation? As mentioned earlier, research on the former topic is still scarce. More empirical work on these topics will provide insights into the washback effects of technology-mediated tests and help language teachers make decisions about what type of assessment works best for their students' learning. Third, research should be done on the issue of at-home test security. As more tests are delivered remotely, especially in response to the COVID-19 pandemic, this line of research will be useful to stakeholders, such as university or academic program administrators who need to make decisions about which language test scores they should accept as evidence of students' foreign language proficiency.

Technology continues to hold tremendous potential for the field of language assessment. It enables and encourages on-going innovation as we harness technology for assessment purposes. The rate of technological development and innovation promises new heights in terms of accuracy, score reporting and affordability that should translate to making tests more accessible to a wider audience. The continued consideration of real-life contexts in which our test takers live and work, including the classroom setting not only in the developed and high-resourced countries, but also in developing countries in low-resource settings remains key to help create opportunities for all test takers. It will remain interesting to watch the continued transition of many areas of work from face-to-face interactions to technology-mediated interactions. The 2020 pandemic has accelerated this development and will likely lead to permanent changes, which in turn will impact assessment practices to align with newly defined real-world tasks. Technology can thus be used to help level the playing field in terms of readily providing authentic tasks to a global audience at a reasonable cost.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments.

Further Readings

Chapelle, C., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press.
This book discusses the theoretical, methodological, and practical issues in computer assisted language assessment and their implications for language teachers and language assessment researchers.

Jones, N., & Saville, N. (2016). *Learning oriented assessment: An overview*. Cambridge University Press.
This book explains what learning oriented assessment means and suggests how language test developers and educators can make it happen.

Zechner, K., & Evanini, K. (2020). *Automated speaking assessment: Using language technologies to score spontaneous speech.* Routledge.
This volume provides a comprehensive overview of state-of-the-art automated speech scoring technology used at ETS and includes topics relevant to ASR systems such as challenges, psychometrics considerations, integral main components, and future directions.

References

Attali, Y. (2004). Exploring the feedback and revision features of *Criterion*. *Journal of Second Language Writing*, *14*, 191–205.

Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater v.2*. (ETS Research Rep. No. RR-04-45). Educational Testing Service.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing*, *31*(2), 241–259. https://doi.org/10.1177/0265532213509810

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355–377. https://doi.org/10.1177/0265532210364404

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the *TOEFL iBT*® test: A lexico-grammatical analysis. *ETS Research Report Series*. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing*, *36*, 3–18. https://doi.org/10.1016/j.asw.2018.02.003

Burstein, J., Elliot, N., Klebanov, B. B., Madnani, N., Napolitano, D., Schwartz, M., Houghton, P., & Molloy, H. (2018). Writing Mentor™: Writing progress using self-regulated writing support. *Journal of Writing Analytics*, *2*, 285–313. Retrieved December 3, 2019 from https://wac.colostate.edu/docs/jwa/vol2/bursteinetal.pdf.

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

Chapelle, C. A., & Voss, E. (2014). 20 years of technology and language assessment in Language Learning & Technology. *Language Learning & Technology*, *20*(2), 116–128. http://dx.doi.org/10125/44464

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. http://dx.doi.org/10125/44145

Chen, L., Zechner, K., Yoon, S., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W. & Gyawali, B. (2018), Automated scoring of nonnative speech using the *SpeechRater*℠ v. 5.0 Engine. *ETS Research Report Series*, 1–31. https://doi.org/10.1002/ets2.12198

Dann, R. (2014) Assessment *as* learning: Blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 149–166, https://doi.org/10.1080/0969594X.2014.898128

Douglas, D. (2013). ESP and assessment. In B. Paltridge & S. Starfield. (Eds.), *The handbook of English for specific purposes* (pp. 367–383). John Wiley & Sons.

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics*, *27*, 115–132. http://dx.doi.org/10.1017/S0267190508070062

Foltz, P. W., Streeter, L. A., Lochbaum, K. E. , & Landuer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis and J. Burnstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 68–88). Routledge

Forsyth, C. M., Luce, C., Zapata-Rivera, D., Jackson, G. T., Evanini, K., & So, Y. (2019) Evaluating English language learners' conversations: Man vs. machine. *Computer Assisted Language Learning*, *32*(4), 398–417, https://doi.org/10.1080/09588221.2018.1517126

Heffernan and Otoshi (2015). Comparing the pedagogical benefits of both Criterion and teacher feedback on Japanese students' EFL writing. JALTCALL Journal, 11(1), 63–76. doi:10.29140/jaltcall.v11n1.184

Jones, N., & Saville, N. (2016). *Learning oriented assessment: An overview*. Cambridge University Press.

Kiddle, T., & Kormos, J. (2011). The effect of mode of response on a semidirect test of oral proficiency. *Language Assessment Quarterly*, *8*(4), 342–360, https://doii.org/10.1080/15434303.2011.613503

Kim, J., & Craig, D. A. (2012). Validation of a videoconferenced speaking test. *Computer Assisted Language Learning*, *25*(3), 257–275, https://doi.org/10.1080/09588221.2011.649482

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*(1), 26–43. https://doi.org/10.1016/j.asw.2007.04.001

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*(3), 319–340. https://doi.org/10.1177/0265532215587391

Lin, T.-J., & Lan, Y.-J. (2015). Language learning in virtual reality environments: Past, present, and future. *Educational Technology & Society*, *18*(4), 486–497. http://www.jstor.org/stable/jeductechsoci.18.4.486

Louma, S. (2004). *Assessing speaking*. Cambridge University Press.

Madsen, H. (1991). Computer-adaptive testing of listening and reading comprehension. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 237–257). Newbury House.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly*, *14*(1), 1–18. http://dx.doi.org/10.1080/15434303.2016.1263637

Ockey, G. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24(4)*, 517–537. http://dx.doi.org/10.1177/0265532207080771

Ockey, G. J. (2018). Oral language proficiency tests. In J. L. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching* (Vol. 3). Wiley-Blackwell.

Park, M. (2018). Innovative Assessment of aviation English in a virtual world: Windows into cognitive and metacognitive strategies. *ReCALL Journal*, *30*(2), 196–213.http://dx.doi.org/10.1017/S0958344017000362

Ramanarayanan, V., Pautler, D., Lange, P., & Suendermann-Oeft, D. (2018). *Interview with an avatar: A real-time cloud-based virtual dialog agent for educational and job training applications* (Research Memorandum No. RM-18-02). Educational Testing Service.

Rudner, L. M., Garcia, V., Welch, C. (2006). An Evaluation of IntelliMetric™ Essay Scoring System. *The Journal of Technology, Learning and Assessment*, *4*(4). Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1651

Salkind, N. J. (2010). *Encyclopedia of research design*. SAGE Publications, Inc. https://doi.org/10.4135/9781412961288

Schultz, 2013. The IntelliMetric™ automated essay scoring engine – A review and an application to Chinese essay scoring. In M. D. Shermis and J. Burnstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 89–98). Routledge.

Shohamy, E. (2000). The relationship between second language testing and second language acquisition revisited. *System*, *28*(4), 541–554. http://dx.doi.org/10.1016/S0346-251X(00)00037-3

Stansfield, C. W. (Ed.). (1986). Technology and language testing. Washington, DC: Teachers of English to Speakers of Other Languages.

Stevenson, J., & Gross, S. (1991). Use of a computerized adaptive testing model for ESOL/bilingual entry/exit decision making. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 223–235). Newbury House.

Tarighat, S., & Khodabakhsh, S. (2016). Mobile-Assisted Language Assessment: Assessing speaking. *Computers in Human Behavior*, *64*, 409–413. https://doi.org/10.1016/j.chb.2016.07.014

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

Turner, C. E. (2013). Mixed methods research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1403–1417). Wiley-Blackwell.

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in the classroom. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 195–208). DeGruyter Mouton.

Vantage Learning. (n.d.) How does IntelliMetric™ score essay responses? Retrieved on November 27, 2019 from http://www.vantage.com/pdfs/research/RB929.pdf

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, *5*(3), 218–243. https://doi.org/10.1080/15434300802213015

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*(4), 493–513.http://dx.doi.org/10.1177/0265532209355668

Wagner, E. (2020). Duolingo English Test, revised version July 2019. *Language Assessment Quarterly*, *17*(3), 300–315, https://doi.org/10.1080/15434303.2020.1771343

Yang, Y., & Qian, D. D. (2020). Promoting L2 English learners' reading proficiency through computerized dynamic assessment. *Computer Assisted Language Learning*, *33*(5–6), 628–652, https://doi.org/10.1080/09588221.2019.1585882

Zechner, K., & Evanini, K. (2020). *Automated speaking assessment: Using language technologies to score spontaneous speech.* Routledge.