Assessment is a critical aspect of any English for academic purposes (EAP) program. In this chapter we start by distinguishing between testing and assessment, and discuss major purposes for assessment. We review the history of language assessment and EAP testing specifically, as well as current critical issues and topics in EAP assessment. Of particular importance in this section is the very definition of EAP. We then examine the current contributions of research to assessing EAP, and the research methods used to investigate it. Based on this review, we make recommendations for practice. Finally, we discuss future directions in testing, particularly with regard to the use of technology in test design and scoring.

We begin our consideration of EAP assessment by defining some critical terms. One area where clarification is needed is the frequent confusion between the terms “assessment” and “testing.” As Bachman and Palmer (2010, p 19) point out, these terms are often used interchangeably to describe a process by which information is gathered in order to make decisions about teaching and learning. For the sake of simplicity, in this chapter we use the term “assessment” primarily to refer to low-stakes, classroom-based assessments, and “tests” to refer to large-scale, standardized assessments that are used to make high-stakes decisions for teaching and learning such as university entrance or school accountability. By this definition, assessments take place in the classroom and are developed by teachers, and tests are developed by professionals and administered to large numbers of students on a school, district, state, national, or international level. In determining whether an information-gathering activity on student performance is a test or an assessment, we must start with the purpose and stakes of the activity. We have said that assessment tends to take place in a classroom and is conducted by an instructor, with the results informing the classroom. These results may be used summatively (Wiliam and Black, 1996) to assess students’ achievement as part of a grade, for example; or formatively to give information about the success or failure of what has been taught and learned (or not) in a specific course, or even whether the instructor should intervene to provide more support for a specific student. By contrast, tests tend to have a broader purpose with higher stakes. For example, a national, state, or district-wide test may provide information on not only the extent to which individual students are meeting the goals of a course but also to what extent individual teachers and schools are successfully teaching the intended material. Similarly, a test that students must pass to meet high school graduation or university entrance requirements is administered on a large scale, and has
Assessment of English for academic purposes

serious ramifications for the students’ future. Tests such as these are considered high stakes because the consequences for the student (in the case of graduation tests) and the teacher (in the case of district or state-wide tests) are high and surpass that of an individual course grade.

In addition to the stakes of a test or assessment, it is important to consider test purpose. Two of the most important test purposes for EAP are proficiency and achievement. A proficiency test is not tied to a particular curriculum but is intended to provide global information about language ability. Good examples of EAP proficiency tests are the Test of English as a Foreign Language (TOEFL) and the International English Language Testing Service (IELTS), which are used to determine a student’s readiness for participating in academic settings. By contrast, assessments used in classes to determine the extent to which specific academic language or content goals have been attained are considered achievement tests. Two additional test purposes important for assessing English for academic purposes are placement testing and diagnostic testing. Placement tests are typically used within institutions to determine what level of English instruction is appropriate for students, and diagnostic tests are used to provide information about specific skills that students should work on. An example of a placement test is the English Placement Test given at the University of Illinois (www.linguistics.illinois.edu/students/placement/); a well-known diagnostic test of EAP is the Diagnostic English Language Needs Assessment (DELNA) developed by the University of Auckland in New Zealand (www.delna.auckland.ac.nz/en.html). Finally, we need to consider what makes a test academic rather than general in nature. A general proficiency test targets for measurement a student’s ability to use the language for everyday purposes; test tasks draw from social and nonacademic contexts such as reading a newspaper comic strip or speaking with a sales clerk. An academic test seeks to simulate as much as possible an academic context; thus, it is crucial to ensure that the texts and test tasks in the assessment are academic in nature. Students in an academic English language class focused on reading will need to be assessed on their ability to read academic texts rather than general, everyday texts such as novels or magazines. Similarly, academic tests of listening should focus on listening to lectures or classroom discussions rather than, for example, announcements in airports. In summary, the assessment of English for academic purposes can range from classroom achievement tests to large-scale international proficiency tests. As with all tests, the test purpose and stakes need to be considered when designing or adopting an EAP test. What distinguishes EAP tests from more general tests of language includes the content or topics, the nature of the language used in the assessment, and the nature of the test tasks, which should all reflect academic settings. We discuss these issues in more depth below.

Historical overview

We begin our review of language assessment in general, and EAP assessment in particular, in the mid-twentieth century. In the 1950s and 1960s, especially in the US, the field of language testing was heavily influenced by measurement theory, with a strong emphasis on reliability of scoring and the minimization of measurement error (Behizadeh & Engelhard, 2011). Influential scholars in language testing at the time included Lado (1961), who advocated using discrete items to test language knowledge, and Carroll (1961), whose two-way matrix of aspects of language (orthography/phonology, morphology, syntax, lexis) and skills (reading, writing, speaking, listening) dominated language testing for many years. Objective tests, predominantly multiple-choice, were consistent with the dominant pedagogical tradition of grammar-translation or audio-lingual methods for second language teaching. The first major international large-scale test of English for academic purposes, the TOEFL, was introduced
by the Educational Testing Service (ETS) during this time, consisting of multiple-choice tests of vocabulary, reading comprehension, listening comprehension, and knowledge of correct English structure and grammar (ETS, 2011).

Starting in the 1970s, trends in language teaching towards more communicative language teaching (CLT) led to parallel trends in assessment. In second language pedagogy, the notion of CLT shifted the goals of language instruction from mastery of discrete forms and rules to developing communicative ability in the second language. This movement led to ideas about communicative language testing and influential models of communicative competence such as those of Canale and Swain (1980) and Bachman (1990). While multiple-choice testing still predominated in the US, there was a shift from discrete testing to more integrated test formats, such as cloze and dictation (e.g., Oller, 1971). During this time, ETS expanded its suite of EAP tests to include the Test of Spoken English (TSE), designed primarily for students wishing to become graduate teaching assistants in US universities, and the Test of Written English (TWE), consisting of a single essay written in response to a brief prompt (ETS, 2011).

One important component of a focus on communicative language in language testing was a new emphasis on analyzing real-word language use situations that examinees would be likely to encounter in school and other language use contexts. In terms of testing for academic purposes, this focus led to careful study of the academic tasks that university students faced in a variety of disciplines (e.g., Ginther & Grant, 1996; Kroll, 1979).

The 1980s brought the development of another large-scale international EAP test for prospective university students, the IELTS. This test was produced jointly by the University of Cambridge Local Examinations Syndicate (UCLES, now known simply as Cambridge English), the British Council, and the Australian International Development Programme (IDP). This replaced the short-lived ELTS, a test aiming at more discipline-specific academic language (with alternative versions for life science, physical sciences, humanities, etc.) which had been developed a decade earlier but proved impractical to administer on a very large scale (IELTS, n.d.).

The increase in the number of international students at universities, combined with the movement towards increasing access to universities to a broader population, led to the in-house development of placement and proficiency tests of academic skills for native and non-native speakers alike, particularly in the area of writing, and in some cases, reading. During this period, more composition and L2 writing specialists advocated successfully to replace indirect tests of writing with direct ones: typically, timed essays based on a prompt and scored on a holistic scale. White’s (1985) influential book on writing assessment argued persuasively against indirect tests of writing, and promoted holistic scoring of writing. In L2 composition, interest in scoring compositions was rekindled when Jacobs et al. (1981) produced a widely used guidebook for assessing L2 composition through weighted analytic scales, incorporating information from L1 composition assessment practices but tailoring it to an L2 audience concerned with specifying linguistic aspects of writing more explicitly.

While the IELTS included a speaking portion from its inception, in the US, the testing of speaking for academic purposes was primarily restricted to the certification of proficiency for teaching assistantships. However, since the introduction of the TOEFL internet-based test (TOEFL iBT®) in 2005, speaking and writing are now explicitly tested on the TOEFL.

Up to this point we have considered EAP testing in terms of post-secondary education. However, in the past 20 years there has been increasing concern about EAP testing for younger learners. In the US, the passage of the federal No Child Left Behind (NCLB) law in 2001 has had a large influence in this area, because all US states are now required to
Annually administer a standards-based assessment of English language proficiency for all English language learners. This law has resulted in the development of new tests of English for academic purposes such as the ACCESS test, developed by a consortium of 33 states (www.wida.us/aboutUs/mission.aspx). Development of the ACCESS and other similar tests has engendered new research into classroom language use in K-12 settings.

A recent trend of importance to EAP assessment is the development of natural language processing and corpus analysis tools. Increased use of these tools has led to two developments relevant to second language assessment. First, corpus linguistics has allowed for a greater understanding of the features of different genres of speaking and writing (see, for example, Biber, 1988; Biber, Connor & Upton, 2007; Swales, 1990, 2004). Large corpora such as the Michigan Corpus of Academic Spoken English (MICASE) and the International Corpus of Learner English (ICLE) have provided opportunities to research various aspects of language use such as vocabulary, syntactic complexity, and discourse moves. This research can inform both instruction and assessment by providing empirical evidence of how successful speakers and writers use linguistic resources in different genres and on different task types. Second, a natural extension of these automated tools has been the development of automated tools for scoring and providing feedback on speaking and writing (see Ware, 2011, for a discussion of automated feedback for L2 writers and Weigle, 2012, for a discussion of automated scoring). Some tests, notably Pearson’s PTE Academic test (http://pearsonpte.com/), are now exclusively computer delivered and scored.

Critical issues

Perhaps one of the most critical issues in EAP assessment is how to define and describe the language to be tested, i.e., the construct. Here we follow the classic definition of Chamot and O’Malley (1994) as “the language used by teachers and students for the purpose of acquiring new knowledge and skills.” (p 40). One implication of this definition is that the demands for academic language differ greatly depending on the academic setting, including student age, purpose of academic study, and demands that are specific to a particular academic context. Logically, the academic language needed for kindergarten students is vastly different than that needed for graduate studies. It is important to note, however, that, in both situations, students need to be able to both understand and use the language of that specific academic setting. Thus, academic language must be defined with a particular setting in mind; the more specific, the more specialized the language, and the fewer (if any) large-scale tests available to assess language for that setting.

In defining the domain of academic language, there are several other considerations as well. First, the use of academic language requires a strong awareness of sociolinguistic appropriateness across the settings in which it is used, from the classroom and interactions with colleagues to professional email exchanges and academic writing. Second, academic language consists of both spoken and written language, and students generally need control of both types of language to excel in an academic environment. Finally, the work of Biber and his colleagues (e.g., Biber et al., 2004, Biber, 2006) in collecting and analyzing academic language in university settings has demonstrated that academic language demands at the university level encompass both specific and general academic contexts. Specific academic contexts include courses and course materials, including face-to-face and online instruction, and reading materials and writing assignments for such courses. In terms of general contexts, students must be able to read, on paper and electronically, a variety of texts, from syllabi and directions for assignments, beyond the technical reading for any course.
Defining the domain of academic language use for younger learners requires a slightly different set of considerations. Research underlying the development of EAP tests for younger English language learners (ELLs) suggests that a critical area of academic language use is the ability of ELLs to understand the language of their instructors in academic settings. Valdes (2004) proposes that the development of academic language stems from the modeling of the different genres of academic language for students by their teachers; the lack of such modeling may have detrimental results for students’ academic language ability. Furthermore, if mandated state tests presuppose student exposure to such language, students without extensive exposure to this language will not be able to perform well. Thus, the academic language modeled to students by their teachers is considered a critical area of academic language, perhaps more so at this level than at the post-secondary level.

The evolving construct of academic language presents several challenges for EAP testing. First, EAP tests measure only a portion of the academic language that students need to develop to succeed in an academic environment. Necessarily, a standardized test of any kind cannot possibly mimic the idiosyncratic linguistic requirements of any particular academic setting, nor can it reflect the interaction inherent in spoken and written academic environments. Second, such tests do not measure subject area content, and therefore do not necessarily reflect a student’s ability to succeed in specific subject areas; that is, a potential student of chemistry may have the general academic speaking and reading ability to survive in a general academic environment but may lack the specific vocabulary and register for chemistry reading and discourse. Conversely, a student well versed in his or her specific academic area nonetheless needs the general academic language to navigate the academic environment beyond classroom content.

**Current trends**

Advances in our understanding of the nature of academic language along with developments in technology have led to new developments in how EAP assessments are created and delivered. This section focuses on two aspects of these developments: computer-based tests and the assessment of integrated skills.

Computer-mediated testing can provide faster results for students and other stakeholders and also reflects the current technology-driven academic environment for many university students, where online learning practices supplement and sometimes supplant face-to-face courses. At the same time, the development of such tests represents a great deal of effort on the test developer’s part, because a computer-based test does not typically mean a direct transfer from paper and pencil to computer. Rather, such changes have required extensive research, including tryouts with students from the targeted population, to ensure not only that the test can be delivered in the computer format (feasibility) but also that students are tested on the same construct and receive similar scores as on a paper and pencil test (validity and reliability).

Since 2000, most international language tests used for admission to English medium universities (TOEFL, IELTS, Pearson) have transitioned, in part or in whole, to computer-mediated testing. The TOEFL iBT® is administered entirely on the computer, with a mixture of human and automated scoring, and the test tasks are quite different from the paper-based TOEFL. The IELTS, in contrast, is currently delivered by paper or computer; both versions are identical in formt, and candidates are allowed to write their essays on paper or computer. The speaking portion of the IELTS continues to be a face-to-face interview (Green and Maycock, 2004). As noted above, the Pearson test is fully automated in
administration and scoring. K-12 English language tests in the US are also beginning a shift to computer-mediated administration. The ACCESS test mentioned previously is currently being transformed from a paper and pencil test (for reading, listening, and writing) and a face-to-face, one-on-one interview between a student and a test administrator (for speaking) to a four-skills computer-delivered test. The goal of this transformation is not to transfer the test directly from paper to computer, but rather to harness new technology to deliver a computer-based test that is more authentic, engaging, and administratively efficient than the paper-based test.

A second trend in EAP assessment is skills integration. Whereas in a test of general English ability for low-proficiency learners it is reasonable to test discrete skills (reading, listening, speaking, writing) separately, the case for separating skills is more difficult to make in academic settings because of the inherently integrated nature of academic language tasks. For example, students at a university are expected to attend lectures, read and discuss textbooks or journal articles, and then write papers using the information gleaned through the other skills. It may make more sense to think of “academic literacy” rather than “academic writing” in recognition of the integration of skills in academic settings. Large-scale tests like TOEFL now include integrated tasks where speaking or writing are based on short listening and/or reading passages, along with independent speaking and writing. Integrated writing tasks are also found in other international EAP tests, including the Canadian Academic English Language (CAEL) assessment and the DELNA in New Zealand (Cumming, 2014).

While integrated tasks are more authentic in terms of their relationship to academic tasks, they are not without their challenges. As Cumming (2014) notes, these challenges include developing high-quality integrated prompts that are easy for students to interpret, along with scoring reliability and rating. In particular, a task that integrates both reading and writing should be scored on a detailed rubric that outlines precisely what reading and writing skills are integral to success on the task. Cumming also notes that validity evidence needs to be gathered to ensure that a test using integrated tasks is measuring the intended skill(s), which may be a more complex endeavor than collecting evidence for tests of discrete skills.

**Main research methods**

A recent comprehensive overview of research methods in language testing can be found in Lumley and Brown (2005), who point out that, while quantitative methods have traditionally predominated in testing research, the past two decades have seen an increase in qualitative approaches to research as well. Quantitative research has been used to investigate both test reliability and test validity, in terms of the internal structure of tests and the validity of inferences based on test scores. The predominant quantitative methods can be generally divided into those that measure similarities between sets of scores (e.g., correlation, regression analysis), and those that measure differences (e.g., analysis of variance). Methods that look at similarities are used to address questions such as the degree to which one (often shorter, less expensive) can predict scores on another (possibly longer, more expensive) test; the degree to which test scores accurately predict success in future academic endeavors; or the degree to which different raters agree on scores of speaking or writing. The other major family of quantitative measures looks at group differences, addressing questions such as whether a test is biased for or against a certain group of test takers, whether tests can detect growth in language learning before and after an instructional period, or whether different test task types elicit different linguistic behaviors which in turn receive different scores.
Quantitative methods are frequently complemented by qualitative methods as the concerns of language testing specialists have broadened to include considerations of both process and product in language testing, and the impact of tests on individuals and societies. Lumley and Brown (2005) mention three main qualitative approaches that are important in language testing: discourse analysis, introspection (verbal report analysis), and ethnographic methods. We discuss these briefly here and then discuss how both quantitative and qualitative methods are used in test validation research.

Discourse analysis refers to a variety of techniques used to analyze and understand test discourse. An important example of discourse analysis in language testing is the role that discourse analytic techniques have played in revealing that the discourse produced in a one-on-one oral interview is qualitatively different from ordinary conversation, thus raising important validity questions about the use of an oral interview to predict performance in other types of interactions (e.g., Johnson, 2001; Lazaraton, 1992). Discourse analysis has also been used recently to investigate validity questions related to writing assessment as well. For example, Cumming et al. (2005) analyzed differences between independent and integrated (source-based) writing tasks, and Knoch and Elder (2010) conducted a discourse analysis of a diagnostic test of academic writing proficiency to investigate differences in the quality of student essays written under different time constraints.

Introspective methods refer to research that involves verbal self-reporting of cognition, primarily of test takers or test raters. Concurrent or retrospective verbal reports have been used to investigate the strategies used by test takers to answer various item types (Cohen, 1984, 1994; Gruba, 1999) and, increasingly, the strategies used by raters in evaluating writing (Cumming, 1990; Lumley, 2002; Weigle, 1994) or speaking (Brown, 2000).

Finally, ethnographic methods, which are meant to provide rich descriptive data of a particular context, include observations, interviews, and questionnaires (Lumley & Brown, 2005). These methods are particularly important in conducting needs analysis when designing tests, and, increasingly, in investigating the uses and consequences of tests in specific contexts. For example, Li (1990) studied the impact of a new test of English for university entrance on teaching in secondary schools.

We now turn to a discussion of how these various methods, both quantitative and qualitative, are used to provide validity evidence for tests. Cumming (2013) lists five areas in which validation research is important: content relevance, construct representation, consistency among measures, criterion-related measures, and appropriate consequences. In this section we define each area and discuss the most commonly used or recommended methods for each.

**Content relevance**

For a test of academic purposes, the test content should be reflective of the academic domain to be tested. Evidence of content relevance comes from either an analysis of the target domain (in this case, the academic setting) when designing tests, or an analysis of test content when adopting an existing test. Domain analysis research can involve such things as corpus analysis of spoken and/or written texts (e.g., Biber, 2006), observations of classroom interaction, task analysis, or other techniques in needs analysis (cf Bocanegra-Valle, this volume). Test content analysis typically involves expert judgments of test content but can also involve gathering information from test takers through interviews or focus groups.
Construct representation

Test tasks should measure the intended construct, and test results should reflect the construct and not unrelated factors. Qualitative research methods such as observations and surveys, analyses of test-taker responses, and think-aloud protocols of both test takers and raters can be used to investigate how well the test tasks elicit particular behaviors, and whether the scores reflect these behaviors as intended. Quantitative analysis of test scores can involve investigating whether different subgroups of learners score in different ways that are not related to the construct of interest (e.g., males vs. females).

Consistency among measures

It is important to ensure that tests are reliable; that is, that the test taker receives essentially the same score regardless of variations in the testing context. Important variables, particularly in a writing assessment, include task and scoring. If different prompts or test tasks are used for different test takers, or if test takers are allowed to choose what topic to write on, steps must be taken to ensure that the tasks or topics are equivalent in difficulty so that scores are not dependent on the particular task. Similarly, since human judgment is involved in scoring (except in rare cases in which scores are generated automatically via computer), it is important to ensure that different raters come to similar scoring decisions. Typically, quantitative methods such as correlations are used to investigate these issues. More sophisticated analyses of the influence of tasks, raters, and other aspects of the testing situation include generalizability theory (e.g., Bachman, Lynch & Mason, 1995; Gebril, 2009) and multi-faceted Rasch measurement (MFRM, e.g., Eckes, 2008). These quantitative measures, which can give an indication of how consistent the measures are, can usefully be complemented by more qualitative research methods such as think-aloud protocols or stimulated recall to investigate the reasons behind any lack of consistency.

Criterion-related measures

One measure of the validity of an assessment is the degree to which it corresponds to other measures of the same skill. For example, scores on an in-house test of academic writing might be compared to scores on a large-scale standardized test, grades on writing assignments, or teacher judgments of student writing ability. Typically, such studies are quantitative in nature and involve correlations (e.g., Powers et al., 2000; Weigle, 2010).

Appropriate consequences

One important area of validation research involves investigating the use of an assessment and its effects (positive or negative) on different groups of stakeholders. For example, mandatory assessments in schools for accountability purposes may lead to improvements in curriculum in some areas but may also lead to a narrowing of the curriculum in others; this phenomenon is frequently referred to as washback. Scholars in validity research from Messick (1989) to the present day have insisted that investigating test consequences and effects on systems such as schools or governmental policies and individuals (teachers, students, materials writers) is an important component of validity research. Such research is typically more qualitative than quantitative, involving ethnographic methods such as observations, interviews, and the like, rather than statistical analyses.
Recommendations for practice

EAP teachers need to be able to design their own classroom assessments as well as prepare their students for, and interpret the results of, any large-scale tests their students will face, whether they are school-based accountability tests or tests that will make a difference to their future, such as the TOEFL or IELTS. Here we discuss the process of test development and some considerations that need to go into the creation of EAP tests.

As any introductory textbook on assessment states, the test development process involves several stages that are iterative and not linear. Excellent advice for adopting, adapting, or developing tests can be found in Brown (2005), Chapter 2; for designing classroom tests in Brown and Abeywickrama (2010), Chapter 3; and, for developing tests of specific skills or in particular contexts, in chapters in any number of recently published handbooks and encyclopedias on language assessment, notably the *Companion to Language Assessment* (Kunnan, 2014).

The steps in test development for a professional examination include:

(a) determining the test purpose
(b) defining objectives
(c) creating a test plan or test specification
(d) writing items
(e) reviewing and revising items
(f) pre-testing and revising items
(g) assembling a complete test
(h) pilot testing
(i) statistical analysis
(j) creating a plan for administration, scoring, and score reporting
(k) ongoing monitoring of the test.

For classroom tests or assessments, these steps will naturally be abbreviated depending on the stakes of the test and the amount of time and resources available; for example, few teachers have the luxury of being able to pilot test their instruments, but teachers may be able to review each other’s tests before administering them to get feedback on clarity of instructions, content coverage, and so on.

An important issue to consider when designing an English test for academic purposes is the degree to which language and academic skills can be separated. For example, in a test of reading for academic purposes, some of the subskills of reading that are considered important by scholars include reading to learn and reading to synthesize information across multiple texts (see, e.g., Enright et al., 2000): skills that rely on cognitive processes that are not necessarily strictly linguistic. Indeed, at higher levels of second language proficiency, a test may be more of a test of academic skills than of language skills, as native speakers of English vary in their ability to succeed on such tasks. This becomes an issue if L2 speakers are tested and L1 speakers are not, for example in tests for international teaching assistants or writing placement tests.

There is a growing consensus that lack of academic vocabulary is an obstacle to student success (Nagy & Townsend, 2012) both in terms of general vocabulary and discipline-specific vocabulary. However, it is not always clear whether academic vocabulary should be assessed discretely or as part of a broader test of academic language. As Nagy and Townsend (2012) state,
because it is generally agreed that academic language cannot be isolated from the disciplinary content it is used to convey (Lesaux et al., 2010; Schleppegrell, 2004), it is unclear whether the best assessments for academic language interventions are measures of disciplinary knowledge or measures of those components of academic language that can be isolated for testing purposes.

Here again we see the tension between assessing discrete bits of language and assessing language in a specific context, which involves both language and disciplinary knowledge.

The final area that is worth mentioning here is the area of citation practices and the avoidance of plagiarism, which is a high-level academic skill that requires a long time to master. The movement towards source-based writing on tests inevitably brings with it questions of how students incorporate words and ideas from the source texts into their writing, and how different types of source use are valued by raters. One recent study (Weigle & Montee, 2012) suggests that raters have very different perceptions of how to evaluate certain kinds of textual borrowing in a source-based writing test, indicating a need for rater training on how raters should account for episodes of inappropriate borrowing when assigning scores to essays.

Future directions

An important new trend in EAP assessment is the expanding use of technology in assessment. As Douglas (2010) points out, language test developers have always incorporated new technologies into their assessments. With regard to technologies of the early twenty-first century, such as podcasting, Voice/Video over Internet protocols (VoIP), or simulated environments (virtual worlds), the question is not whether these technologies should be used, but rather how to use these technologies in a manner that is both responsible and efficient. Douglas poses the following essential questions:

(a) What are the effects of the uses of technology in language assessment on test taker attitudes such as anxiety and motivation?
(b) How does language performance differ with different technologies?
(c) How does the use of technology affect the definition of the language ability construct we are attempting to measure?
(d) How does technology affect the nature of assessment tasks that can be developed?
(e) What are the limits of technology for scoring constructed response assessment tasks and how does automated scoring differ from human scoring?

(2010, pp. 116–117)

Douglas provides some potential answers to the applications of technology with assessment in general; here we discuss them in terms of their applicability to EAP assessment. First, in terms of motivation, the landscape has changed considerably over the past twenty years, when computers were less likely to be present in homes and classrooms. When the TOEFL went from paper and pencil to computer, for example, scholars were concerned that the computer interface might be unfamiliar and cause undue anxiety to test takers. These days, however, the opposite may be true—students are so used to working on computers that they may feel less motivated to take a paper and pencil test than to write on the computer. However, it is also important to note that technology may both help and confuse instructors, who may need additional technical support to navigate new technologies and, in fact, may in some cases be more reluctant than their students to embrace new approaches.
As for technology’s effects on language performance, research has demonstrated that both reading and writing processes differ between paper and pencil and digital environments; it thus becomes important to specify the role of technology in the language use situation of interest. For example, in a test of English for academic purposes at the secondary or post-secondary level, in settings where students will be expected to access readings online, post responses to electronic discussion boards, or write research papers, the ability to use language through electronic technologies might very well be part of the construct to be measured, and thus important to include in test task design.

The area of automated scoring is one that has generated a good deal of interest in recent years; see, for example, a special issue of Language Testing devoted to automated scoring and feedback systems (Xi, 2010). While new technologies in assessment, such as automatic scoring, can decrease the costs of assessment rating, these technologies add a new layer of responsibility to the testing company to ensure that the ratings produced by automated scoring engines are consistent and as valid and reliable as those of human raters. Similarly, such efforts raise the question of potential student attitude toward a test they know is partially or completely rated by a computer rather than a human being.

In summary, new technology can be a boon to EAP assessment if it allows for innovative tasks that replicate authentic academic tasks; however, care must be taken to ensure that new technologies enhance rather than reduce assessment validity.

**Further reading**

Coombe, O’Sullivan & Stoynoff (2012); Schleppegrell (2004); Shermis & Burstein (2013); Weigle (2002)

**Related chapters**

11 Language and L2 writing
35 The Common Core in the United States
36 EAP pedagogy in undergraduate contexts

**References**


Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 5–43.


Sara Cushing Weigle and Margaret E. Malone


