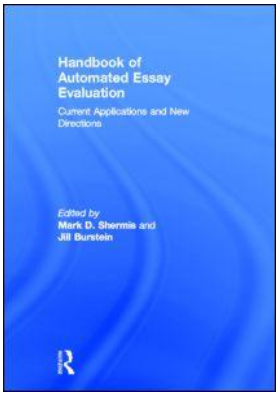

Handbook of Automated Essay Evaluation
Current Applications and New Directions

Mark D. Shermis, Jill Burstein

English as a Second Language Writing and Automated Essay Evaluation

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780203122761.ch3
Sara C. Weigle
Published online on: 18 Apr 2013

3 English as a Second Language Writing and Automated Essay Evaluation

Sara C. Weigle

INTRODUCTION

Much of the published work on automated scoring of writing has focused on writing instruction and assessment in K–16 education in the United States, with the implicit assumption that the writers being assessed have native or native-like command of English. In this context, English language learners (ELLs) appear to be of somewhat peripheral concern. However, readers may be surprised to find out that there are more people learning and communicating in English in the world than there are native speakers (Kachru, 1997; McKay, 2002); some are immigrants or children of immigrants in the United States (U.S.) or other English-speaking countries; some intend to work or study in an English-speaking country; and some use English for professional reasons in their home countries. Given the importance of written communication for global education and business and thus the need to teach and assess writing throughout the world, the interest in reliable and efficient automated scoring systems for assessing the writing of ELLs is increasing, and the applicability of automated essay scoring (AES) systems to non-native speakers (NNSs)[1] of English is an ever-more important concern. One might even argue that the largest market for automated scoring of English writing is not in assessing the writing ability of native speakers, but rather that of NNSs of English.

The goal of this chapter is to lay out some groundwork for understanding the potential place for AES systems for ELLs, whether they are in school with their native speaker peers in the U.S. or learning English as a foreign language (EFL) in a country where English is rarely encountered outside the classroom. I will discuss both AES, defined as "the provision of automated scores derived from mathematical models built on organizational, syntactic, and mechanical aspects of writing" and automated feedback, or "computer tools for writing assistance rather than for writing assessment" (Ware, 2011, p. 769), since these two applications of AES have different considerations for use.

The chapter is organized as follows. First, I discuss the main contexts for assessing writing among ELLs, both in English as a second language (ESL) and EFL settings. Then, I discuss the construct of writing for different populations of ELLs, specifically with regard to the relative importance of second language proficiency and writing ability in different contexts, and current trends in teaching writing to ELLs. Next, I describe briefly the two main functions of computer-assisted writing evaluation: scoring and feedback, and how these functions are implemented in existing systems. Finally, I discuss considerations for implementing automated scoring and automated feedback in different contexts.

CONTEXTS FOR TEACHING AND ASSESSING ELL WRITING

Second language writing instruction and assessment varies widely depending on the context and population of concern. The first major distinction is between ESL and EFL contexts. In the U.S. and other English-speaking countries, there are two main populations of ELLs for whom writing assessment is important. Many NNSs in schools and universities in the U.S. and other English-speaking countries are in writing classes with native speakers, and thus are assessed in the same way, e.g., for placement into composition courses or achievement in class. Even if these students frequently do not have complete control over English vocabulary and grammar, their language proficiency is presumed to be strong enough so that the focus of assessment can be writing *per se*, not language proficiency as demonstrated through writing. The main question for these students is whether assessments designed for native speakers, whether scored by human raters or computers, are valid and fair for NNSs.

The other main writing assessment purpose for ELLs in English-speaking countries is to evaluate English language proficiency through writing. In these settings, there is a greater focus on control over syntax and vocabulary, along with rhetorical concerns such as development and organization. In K–12 and university settings, NNSs are often tested to determine whether they need additional English language services before or concurrent with their regular program of study. In the U.S., the No Child Left Behind law (NCLB) requires ELLs to be tested annually in listening, speaking, reading and writing (No Child Left Behind Act, 2001). For this population, the focus of the assessment tends to be language proficiency as demonstrated through writing, rather than strictly writing skills *per se*, especially at lower levels of proficiency. The question here is whether one or more writing samples can provide sufficient information about a student's language proficiency to make useful decisions about whether he or she needs additional language support to access academic content and perform successfully.

In EFL contexts, there are three main purposes for assessing English writing ability. A large testing industry has grown around the need to assess English language proficiency for students coming to the U.S. or to other English-speaking countries. In the U.S. the main test for this purpose is the Test of English as a Foreign Language (TOEFL); in the U.K. and Australia the most familiar test is the International English Language Testing System (IELTS). More than 27 million people have taken the TOEFL to date (ETS 2012), and 1.7 million take the IELTS (IELTS.org). Secondly, English is a requirement for university students in many countries. For example, in China, approximately 13 million students took the College English Test (CET) in 2006 (Zheng & Cheng, 2008), and South Korea is currently developing its own English language test; both of these major examination systems including writing tests.

Third, there is a great deal of international interest in developing English language tests for workplace certification, particularly tests aligned with the Common European Frame of Reference (CEFR) standards (Europe, 2001) that have been promulgated over the past ten years. Among the most well known of these tests are the Cambridge suite of exams, including the First Certificate in English (FCE) and the Certificate in Advanced English (CAE) (http://www.cambridgeesol.org/index.html).

English examinations have fairly high stakes for students, as their future may depend on their test scores. Thus, test preparation is a large industry in many places, including general English courses or courses specifically tailored towards preparation for a specific exam. To the extent that writing is a central component of the examination, writing will be part of the curriculum. In all of these situations, the primary purpose of these assessments is to evaluate language ability in a specific context—academic or vocational, for example—through writing.

Second Language Writing Ability

The discussion of contexts above highlights the fact that ELLs differ greatly in terms of a number of variables (e.g., age, educational level, ESL vs. EFL, proficiency level) that influence how the construct of writing is defined for a given purpose. Crucially, these variables are related to the degree to which the focus of instruction and assessment is on linguistic or rhetorical concerns; that is, concerns about mastering sentence structure, morphology, and vocabulary vs. concerns about higher-level issues such as audience, voice, and genre. This focus, in turn, has major implications for the use of AES and to its acceptance by stakeholders, including students, teachers, administrators, and users of test scores.

There is a general consensus in the field that second language writing ability is dependent upon both writing ability and second language ability (e.g., Cumming, 1989; Weigle, 2002; Williams, 2005), though the exact contribution of each, and indeed, the degree to which they can be neatly separated, are a matter of dispute. Native English speakers learning to write essays in school generally have automatic control over basic text production processes and extensive experience with English texts; thus, although their academic vocabulary and control over advanced structures such as relative clauses and non-finite subordination strategies may still be growing, they do not need to devote cognitive effort to subtler aspects of English grammar such as article and preposition use or the formation of verb tenses. NNSs of English, on the other hand, vary tremendously in their control over English syntax, morphology, and vocabulary, their familiarity with English written genres, and their experience in writing, either in English or in their home language. Those who are strong writers in their first language can often transfer writing skills to their second language given a certain level of proficiency. However, limited English proficiency can hamper student writing because of the need to focus attention on language rather than content (Weigle, 2005). At lower levels of language proficiency, then, the focus of assessment is generally on linguistic issues; that is, the degree to which writers have control over basic vocabulary, syntax, and paragraph structure. As writers gain more control over these skills, the focus can shift to higher order concerns such as development, strength of argument, and precision in language use. Finally, at the highest levels of proficiency, second language writers may still retain an "accent" in writing but otherwise do not need to be distinguished from first language writers in terms of assessment.

From this discussion it is clear that what is meant by writing in assessment can be very different in different contexts, and that therefore when we talk about automated scoring of writing we need to be clear about what kind of writing is meant. While automated scoring of writing has been quite controversial in the composition community (e.g., Conference on College Composition and Communication, 2004; Crusan, 2010), it may be less controversial when the focus of the assessment is on English language proficiency rather than composition ability (Weigle 2010, 2011). Deane (in press) argues that

> when the use case emphasizes the identification of students who need to improve the fluency, control, and sophistication of their text production processes, or affords such students the opportunity to practice and increase their fluency while also learning strategies that will decrease their cognitive load, the case for AES is relatively strong; but if the focus of assessment is to use quality of argumentation, sensitivity to audience, and other such elements to differentiate among students who have already achieved fundamental control of text production processes, the case for AES is relatively weak.

The former is precisely the case for many ELLs both in ESL and EFL settings, and thus a relatively strong argument can be made for automated scoring and feedback systems if they can be implemented wisely.

Writing Instruction for L2 Learners

Before discussing the use of AES for scoring or feedback on writing, it is important to examine practices in writing instruction for ELLs. As discussed above, writing instruction for L2 learners varies by age, proficiency level, and context. However, some pedagogical principles apply across instructional settings. A good place to start is Silva's (1993) conclusion from an extensive literature review that composing in a second language is generally "more constrained, more difficult, and less effective" than writing in a first language. Depending on how proficient they are, therefore, students learning to write in their second language need more of everything: they need more examples of written texts to learn from, more practice writing, more opportunities to develop effective writing strategies, more familiarity with genres, more practice with vocabulary and grammar, and more feedback. Writing teachers, especially in second language academic contexts such as first-year composition courses, need to find ways to balance the need to provide opportunities to learn and practice new language structures with opportunities to improve written fluency without getting bogged down in grammatical concerns.

It is well known that language proficiency, as it relates to writing, develops slowly over a number of years and depends on extensive exposure to different texts in different genres. Certain elements of grammar, for example, appear to be resistant to explicit instruction and acquired late, such as the use of relative clauses and the English article system (Ellis, 2005). However, other aspects of writing seem to be amenable to instruction regardless of English proficiency level. For example, Roca de Larios, Murphy and Marin (2002) suggest that writing strategies such as problem-solving strategies, goal setting and organization, and having a sense of audience can be effectively taught within the course of a single semester.

The major paradigm in writing instruction in the U.S. for teaching students who are at a stage of language development where they are able to compose texts of a paragraph or longer is the so-called "process approach," also dominant in L1 writing instruction (e.g., Emig, 1971; Flower & Hayes, 1981). That is, the process of idea generation, drafting, giving and receiving feedback, and revising, which is used by expert writers, is modeled and supported by the teacher. In such an approach, students submit multiple drafts of essays and only the second or third draft of the essay is graded. The process approach is contrasted with a "product approach" (e.g., Kroll, 2001) in which students are graded on the basis of a single draft, with no opportunity to revise.

In other countries, of course, writing instruction varies across educational settings. In many places writing is viewed as a support skill for reinforcing and practicing the grammar and vocabulary learned in class, rather than as a skill to be developed in its own right for communication. As many EFL teachers are NNSs themselves, they often do not feel equipped to write or comment on student writing in English, and much of the writing that is done in class is in preparation for examinations. As a result there are many settings where a process approach is not implemented; rather, the focus may be explicitly on practicing strategies for writing timed essays of the kind that are typical on large-scale assessments (see, for example, You, 2004).

The Role of Error Correction in Second Language Writing Instruction

One of the features of AES systems emphasized by proponents is the ability to provide instantaneous feedback on writing, and particularly on sentence-level errors in grammar and usage. Thus, it is important to explore the role of error correction in L2 writing instruction. After all, if error correction is not useful and does not lead to improvements in writing, as claimed by Truscott (1996, 1999, 2007) and others, there is little point in automating it.

In the U.S., the process movement for some time de-emphasized the language aspect of writing instruction for L2 learners. However, the pendulum has swung back the other direction with the understanding that second language writers need and want to improve their knowledge of academic English grammar and vocabulary. One implication of this is that students want feedback on their writing, particularly about the errors that they make. The issue of error correction is one that has caused a great deal of controversy and anxiety for second language writing teachers, particularly in light of some scholarship that suggests that error correction is unhelpful or even harmful (e.g. Truscott, 1996, 1999, 2007). It is clear from the research that L2 writers want comprehensive error correction; it is also clear that teachers frequently do not always know how to provide useful feedback on errors or feel that the time spent in providing comprehensive error feedback reaps useful benefits.

Ferris (2011) provides a thorough review of the research on corrective feedback in L2 writing. This research suggests that students welcome feedback on errors, and that focused feedback which can be used in revising has both short-term and longer-term effects on their control over specific structures targeted in feedback. In particular, Ferris notes that students feel that "teacher feedback on grammar and errors is extremely important to their progress as writers" (p. 46); they prefer comprehensive marking of errors and strategies for correcting them rather than direct correction; on the other hand, they often find teachers' marking systems confusing.

The degree to which L2 teachers actually provide accurate and comprehensive feedback is unclear. Although an often-cited paper by Zamel (1985) reports that teachers are inconsistent, arbitrary, and contradictory in their comments, there are no statistics in the paper to substantiate this assertion, and it is quite likely that 25 years of improved teacher education has improved the situation substantially. In a study designed to gather data on the nature and effect of teacher feedback, Ferris (2006) collected writing samples from six sections of an ESL composition course containing teacher feedback on 15 error types (e.g., word choice, verb tense, word form, articles) using a coding system that the instructors had agreed on. Ferris found that teachers correctly marked 89% of errors identified by independent researchers in one sample, and 83% in another sample. While these results cannot be generalized to other settings, they at least provide a baseline on which to judge automated scoring systems: if 80–90% agreement is a reasonable outcome for experienced teachers working within the same program, one could argue that this would be a reasonable ultimate goal for automated scoring engines to achieve.

To summarize, in some contexts writing instruction for ELLs is very much focused on language acquisition, while in others such instruction is focused on writing strategies. In reality, most ELLs need both types of instruction; one complication is that teachers dealing with ELLs are often trained either in second language acquisition or in composition pedagogy, but frequently not both. For example, in many MA TESOL programs, a course in teaching writing is not required; on the other hand, graduate students from English departments in U.S. universities teaching first-year composition frequently have second language writers in their courses but do not necessarily have specific ESL train-

ing. In order to understand an appropriate role for AES in L2 writing assessment, it is important to take what we know from both perspectives.

Specifically, for automated scoring, it is important that scoring systems be able to identify the features of language that characterize learners at different proficiency levels (and not just errors), and, at the same time, allow learners who are still developing in their language to demonstrate their writing competence (overall content development, quality of ideas, organization, etc.) without being penalized for errors that do not interfere with comprehension or that are late acquired and not amenable to instruction. This is particularly important when ELLs are assessed along with their native speaker peers.

In terms of automated feedback, ELLs want and need specific feedback on language, but such feedback must be digestible and written in a way that is useful for learning. ELLs also want and need feedback on content and organization, but features associated with these aspects of writing are more challenging from a computational perspective to provide automatically. As Chapelle and Chung (2010) have suggested, the ideal hybrid may be one in which sentence-level feedback is provided by an automated mechanism, leaving the teacher free to comment on higher order concerns. Nevertheless, research in natural language processing is focusing on these higher-order concerns, as discussed in later chapters: Automated Short-Answer Scoring (at Educational Testing Service); Automated Evaluation of Discourse Coherence Quality; and, Automated Sentiment Analysis for Essay Evaluation.

AUTOMATED ESSAY EVALUATION: SCORING AND FEEDBACK

The difference between automated scoring and automated feedback is important when considering their use for ELLs. Automated scoring is primarily intended for large-scale tests (that is, beyond the level of the classroom) and automated feedback is primarily intended for instructional use, although these two functions are sometimes blurred. For instance, it is possible, and may be very welcome, to provide diagnostic feedback on essays in large-scale tests. It is also possible, though perhaps less desirable, to incorporate automated scoring along with automatic feedback in a classroom setting. For the moment, however, we will make a distinction between scoring—that is, using automated tools to produce a score that is intended to be equivalent to a human score on the same essay for the purpose of some decision, such as admission or placement—and feedback: the use of automated tools to provide information that will help students improve their writing. As noted above, automated scoring for high-stakes decisions is a highly controversial issue within mainstream composition studies. Automated feedback, on the other hand, is viewed somewhat more positively as a supplement to teacher feedback in classroom use (Ware, 2011; Warschauer & Grimes, 2008).

Large-scale tests are often produced by private companies employing testing and measurement experts and are used to make relatively high-stakes decisions, such as admission to higher education, graduation, and placement. Examples of such assessments include the TOEFL, the CET required of all Chinese students graduating from any college in China (Zheng & Cheng, 2008), and the Entry Level Writing Requirement required of all incoming first-year students in the University of California system (California, 2012). Some tests, such as the Graduate Record Examination, are taken by L1 and L2 students without distinction; others are designed specifically as English proficiency tests for L2 speakers; and yet others, such as the University of California test, incorporate mechanisms for determining whether students who need writing courses would benefit from classes specifically designed for L2 speakers. Some of these assessments currently use

AES systems; for example, e-rater® has been used along with human raters on both the TOEFL® and the GRE® (Educational Testing Service, 2012). In other cases, AES is not currently used but may be under consideration or development for the future.

The most obvious potential advantage of AES for large-scale assessment is the savings in terms of time and cost, given the labor-intensive nature of human scoring of writing, as well as the reliability of AES in producing the same score for a given essay. Most large-scale tests require essays to be double-rated for reliability, with a third rater used if the first two raters disagree. At a single university with a test for only a few hundred students, scoring essays can require several raters working over multiple days, and many tests are much larger. An automated scoring system that can accomplish this scoring in a matter of minutes represents a major savings of time, if not necessarily money.

On the other hand, the major objection to automated scoring for high-stakes exams is the argument that a computer cannot "read" an essay and that the most important features that contribute to essay quality are not quantifiable. Furthermore, the decisions made on the basis of test scores are too important to be made by machines. Other objections have to do with the lack of transparency about the algorithms used to score essays, the de-professionalization of teachers, the narrowing of the construct, and the potential consequences of using computers to score writing (Cheville, 2004; Condon, 2006; Crusan, 2010; Herrington & Moran, 2001). One of the goals of this volume was to address the issue of system transparency. Several systems are described in the chapters that follow.

In contrast to large-scale testing, much writing assessment is done at the classroom level by a teacher who knows the students, is in control of the curriculum, and understands the context for the assessment. Within classroom assessment we can distinguish summative and formative assessment. Summative assessment is used to evaluate how much students have learned in a course and whether specific aims have been met. Formative assessment, on the other hand, is used to help teachers and students diagnose and address specific writing problems during a course, before a final grade is recorded.

Considerations for the use of automated scoring and feedback differ for these different situations. A better argument can be made for AES—both scoring and feedback, but particularly feedback—for formative assessment rather than for summative assessment. In the context of formative assessment, systems that allow students to submit an essay for instant scoring, receive useful feedback, and revise an essay with the goal of getting a better score can be motivating for students (Grimes & Warschauer, 2010). However, in terms of summative assessment, the final determination of what a student has achieved in terms of meeting the instructional goals for a course is surely better left to the judgment of the teacher rather than to an automated system.

In summary, it is possible to make a plausible case for AES in high-stakes large-scale assessment, provided that there are sufficient checks within the system to ensure ongoing accuracy and reliability of scoring and evidence in addition to computer-generated scores on which to base important decisions about a student's writing proficiency such as placement or university admission. It is also possible to make a plausible case for AES in very low-stakes assessment, that is, automated feedback for formative assessment within the classroom. Here, the important checks to be made are the degree to which feedback can be understood and used by students and how well teachers are trained in system use so that they can make use of its advantages appropriately. On the other hand, the case for AES is weakest as a tool for summative assessment in the classroom, where the classroom teacher is the most appropriate person to make the ultimate evaluation as to how much students have learned or achieved. Teachers are rightly concerned about the dangers of handing over critical decision making to a computer that may be focusing on the wrong

skill, and de-professionalizing teaching by using automating grading to make critical decisions about students' education.

Features of AES Systems

Given the potential for using AES systems with ELLs, I now turn to a discussion of what AES systems can and cannot do. Commercially available AES systems, such as e-rater, created by Educational Testing Service (ETS), and IntelliMetric, from Vantage Learning, can produce ratings on a holistic scale that agree with one human rater as often as two human raters agree with each other, and can do so much faster than human beings (Attali & Burstein, 2006; Burstein, 2002; Burstein & Chodorow, 1999; Elliot, 2003; Landauer, Laham, & Foltz, 2003; Page, 2003). Automated scoring systems may use natural language processing or Latent Semantic Analysis techniques that provide measures of quantifiable aspects of essays, which can be combined mathematically to predict human scores with a high degree of accuracy. Readers should refer to the e-rater chapter for a brief discussion of natural language processing methods.

It is important to recognize that the features identified and counted by AES systems are not necessarily the things that human raters pay attention to in rating essays; however, the argument can be made that the quantifiable features stand in as proxies for the features that raters value (see, for example, Connor-Linton, 1993; Cumming, 1990; Huot, 1993; Lumley, 2002; Milanovic, Saville, & Shen, 1996; Vaughan, 1991 for discussions of rater behavior and decision-making processes). For example, e-rater evaluates organization and development by automatically identifying the presence or absence of relevant discourse units (e.g., introduction, thesis statements, main ideas, supporting details, and conclusion). In addition to identifying discourse element types to represent organization, the system also takes into account the length of each element identified to represent development. (See the chapter about the Automated Evaluation of Discourse Coherence Quality for a discussion of natural language processing research in this area.) By comparison, raters pay attention to the organization and flow of an essay, in terms of an introduction, body paragraphs with supporting statements, and a concluding paragraph. The human rater will consider whether the organization is easy to follow, whether it makes logical sense, and whether the conclusion wraps up the essay in a satisfactory way. These things may not be quantifiable, as they rely on the fact that the reader brings background knowledge and expectations to the reading and intuitively compares the essay to others that he or she has read in the past, and often to knowledge about the writers. However, even if neither the human rater nor e-rater explicitly considers essay length as a central indicator of writing quality, the length of discourse units (in the case of e-rater) frequently corresponds with human judgments of development, as it is difficult to develop an essay well without expanding on main points and supporting them with relevant details, all of which have the effect of increasing the length of such units.

Similarly, text analysis research has suggested several features of essays that tend to correlate well with essay scores, ranging from final free modifiers (Nold & Freedman, 1977) to the number of words before the main verb (McNamara, Crossley, & McCarthy, 2010), or error-free T-units (Homburg, 1984). Again, even though raters may not specifically focus on such features, the relationship between these features, which can be counted automatically, and rater scores reflects the fact that writers who are able to express complex ideas generally have mastered the syntactic means to do so in a sophisticated way. Thus, while some of the features of automated scoring systems cannot replicate what human raters respond to, they may stand in as proxies for those constructs.

Another reason why automated scores are generally consistent with human scores has to do with the fact that the features that raters pay attention to in scoring tend to be highly correlated with each other. There are two possible explanations for this: first, the different sub-skills develop more or less in tandem, as the writer gains more experience reading and writing, and second, raters are not necessarily able to distinguish one from another, even when using an analytic scoring rubric (Marsh & Ireland, 1987). For example, the word choices that a writer uses, which is frequently captured in a scoring rubric under the category of language use, can influence the reader's understanding of content or organization.

There certainly are some cases in which an automated scoring system will not give the same score as a trained human rater. For one reason, AES systems cannot "read" a text in the same way that a person does. That is, while it is possible for AES systems to parse sentences and identify propositions, they cannot relate those propositions or texts to a body of world knowledge or judge the reasonableness of certain types of evidence in support of a thesis. Thus, AES systems are not currently well suited to evaluate features of writing such as authorial voice or strength of argument, which depend on shared background knowledge and assumptions between reader and writer. However, as is discussed in the Sentiment Analysis and Discourse Coherence Quality chapters later in this volume, there are current research efforts intended to address features related to both authorial voice, and argumentation.

By the same token, the computer cannot detect features of an essay that might be rewarded by a human rater, such as allusions to well known literature, people, or events that may be unusual but apt in the circumstances, or the use of humor or irony. One example of this can be seen in ETS's own materials. In a screen shot in the online Criterion demo, for example, the sentence "monkey see, monkey do" is marked as having subject/verb agreement errors. From a strictly grammatical perspective this is true, but from a pragmatic perspective the sentence is perfectly well formed, evoking the reader's (stereotyped) world knowledge about monkey behavior and the appropriate use of proverbs in writing, among other things. We have not yet reached a point in automated scoring where the computer can recognize pragmatically skillful uses of 'incorrect' language like this.

Furthermore, the computer cannot easily judge the intended meaning of a writer when errors are made. ESL errors are notoriously difficult to categorize, and AES systems may not able to use context reliably to distinguish among possible interpretations of ill-formed sentences. Chodorow, Gamon and Tetreault (2010, p. 422) give the example "I fond car": does this involve a misspelling of 'found' and a missing article (I found the car) or a missing copula, preposition, and plural marking (I am fond of cars)? In the context of an actual essay a human rater would be able to deduce from the context what the most likely interpretation would be, but AESs are not yet at the point where they can reliably do so. Another aspect of concern is that there are degrees of seriousness of errors, in terms of what errors are stigmatizing and what create confusion. Because it is impossible to predict exactly what errors will lead to difficulties in comprehension, it is difficult for an AES system to predict and evaluate the seriousness of different error types.

Despite these differences, in the majority of cases, automated scoring systems correlate with human judgments about as well as humans do with each other. A few advantages of AES are that they are not affected by factors that make human raters unreliable, such as fatigue, inattention, or distraction. On the other hand, there are still several validity questions that need to be answered regarding automated scoring. Xi (2010) provides a list of questions that should be asked of automated scoring systems, corresponding to the different steps in a validity argument (Chapelle, Enright, & Jamieson, 2008; Kane, 1992, 2006). These questions are listed in Table 3.1.

Table 3.1 Validity Questions Related to Automated Scoring

Inference	*Validity Questions*
Domain representation: the performance represents the target domain	Does the use of assessment tasks constrained by automated scoring technologies lead to construct under- or misrepresentation?
Evaluation: the scores are accurate representations of the performance	Does automated scoring yield scores that are accurate indicators of the quality of a test performance sample? Would examinees' knowledge of the scoring algorithms of an automated scoring system impact the way they interact with the test tasks, thus negatively affecting the accuracy of the scores?
Generalization: the observed score is an appropriate estimate of other scores obtained from other, similar observations	Does automated scoring yield scores that are sufficiently consistent across measurement contexts (e.g., across test forms, across tasks in the same form)?
Explanation: the scores can be attributed to the construct	Do the automated scoring features under- or misrepresent the construct of interest? Is the way the scoring features are combined to generate automated scores consistent with theoretical expectations of the relationships between the scoring features and the construct of interest? Does the use of automated scoring change the meaning and interpretation of scores provided by trained raters?
Extrapolation: the scores are indicative of performance in the target domain	Does automated scoring yield scores that have expected relationships with other test or non-test indicators of the targeted language ability?
Utilization: the scores provide useful information for decisions and curriculum	Do automated scores lead to appropriate score-based decisions? Does the use of automated scoring have a positive impact on examinees' test preparation practices? Does the use of automated scoring have a positive impact on teaching and learning practices?

Source: Xi, 2010.

A complete treatment of these validity questions is beyond the scope of this chapter. However, the literature has paid more attention to some of these questions than to others, and it is to those that I turn my attention now. First is the issue of construct underrepresentation; that is, whether the use of automated scoring constrains the assessment tasks that can be used. Typically, the independent expository or argument/persuasive essay has been the most frequent candidate for automated scoring. However, the types of tasks used in L2 writing assessment are expanding to cover a more broadly conceived notion of writing ability. A recent trend in large-scale tests such as the TOEFL is the use of integrated tasks, in which students receive input through listening and/or reading and use this information in a written response. Such tasks are claimed to be more authentic than traditional independent tasks because academic writing nearly always involves reading as well (Carson, 2001; Feak & Dobson, 1996; Weigle, 2004). It is not clear how automated scoring systems would be used to evaluate integrated responses. One complication is how to deal with the presence of language from the source texts in the written responses, some of which may be appropriately or inappropriately used (see Weigle & Montee, in press, for a discussion of rater perceptions of source text borrowing in integrated tasks). The investment in automated scoring systems to evaluate a specific type of writing may make companies reluctant to experiment with other task types, which could have the effect of narrowing the construct.

A perhaps more serious challenge with regard to construct narrowing comes from scholars such as Condon (2009), who notes that the writing produced in a typical writing test is of very little interest beyond the assessment itself. Condon presents an alternative type of writing assessment task at the university level, which requires students to reflect on and write about what they have learned at the university. Condon claims that the content of the writing is not only more engaging and interesting to the readers, but serves an authentic communicative purpose that extended beyond the specific testing situation. Condon's argument is not necessarily specific to automated scoring; his focus is on rethinking the whole notion of timed essay tests as a way of evaluating writing. In his view, automated scoring serves to reify the essay exam as the main assessment tool, and scholars should be cautious in developing automated scoring systems without critically examining the basis of the assessment.

A related question is whether actual or perceived knowledge of the algorithms will change the way students prepare for an exam. In situations where test scores have very high stakes for students, the curriculum tends to focus on strategies for passing the test. You (2004), for example, describes the English curriculum at a Chinese university, noting that

> a typical college English curriculum in China works under the guidance of the College English Syllabus and is evaluated almost exclusively by the results of students' scores on the CET. In such a curriculum, students' individual needs for English are hardly acknowledged; many teachers are predominantly concerned about teaching language knowledge and test-taking skills, instead of language skills for communication purposes. English writing is still taught in the current-traditional approach, focusing on correct form rather than helping the students develop thoughts. Systematic language instruction is severely constrained by simulation tests and various test-preparation exercise manuals when the CET draws near.
>
> (p. 108)

Thus, the possibility of automated scoring algorithms affecting the way students learn to write is quite real. Both Grimes and Warschauer (2010) and Chen and Cheng (2008) report that teachers and students believed that the scoring algorithm favored the traditional five-paragraph essay and that the scoring algorithms discouraged creativity in writing, suggesting that students wanting to get higher scores may be less likely to take risks in their writing.

One important validity question is whether automated scores and human scores have the same relationship with other indicators of writing ability. This issue was taken up by Weigle (2010, 2011) in the context of the Internet-based (iBT) TOEFL. In the study, relationships between human and automated scores on TOEFL iBT essays, on the one hand, and a variety of indicators of writing ability, including self-assessment, instructor assessment, and ratings of non-test writing samples, on the other, were investigated. Correlations between essay scores (whether generated by human raters or computers) and these indicators were moderate, with higher correlations with more global language proficiency measures such as self-assessment than with more specific writing-related issues. Human and e-rater scores differed in their relationships to indicators of writing ability only in a few cases. Most strikingly, the scores of human raters on TOEFL iBT essays were more strongly correlated with judgments made by content area teachers about the seriousness of language problems faced by participants than were e-rater scores, suggesting that both the human raters and the instructors were more sensitive to some aspects of student writing than was e-rater. Weigle's data supports the use of e-rater with the

TOEFL, bearing in mind that the main use of the TOEFL is to assess the "ability to use and understand English at the university level" (Educational Testing Service, 2012), rather than writing *per se*.

Xi (2010) notes that the higher the stakes, the more validity evidence needs to be presented to justify any use of automated scores, particularly if automated scores are used alone. At present there are no plans to use e-rater on any large-scale assessment without a human rating as well, and the results from Weigle (2011) support the idea that automated scoring should be complemented by human judgments, even if the focus of the assessment is on linguistic rather than rhetorical issues such as argumentation and voice.

Automated Feedback

I now turn to consideration of automated feedback, which holds a promise of reducing teacher's burdens and helping students become more autonomous. AES systems can recognize certain types of errors and offer automated feedback on correcting these errors, in addition to providing global feedback on content and development. One benefit of automated feedback for many students is that it is anonymous and not personal, allowing students to save face in a way that submitting their writing to teachers does not allow. Another benefit is that students can receive instant feedback, repeatedly, on an essay. Teachers may not have time to give detailed feedback on numerous drafts of essays before giving a final grade; on the other hand, students can submit their essays to the computer for scoring as many times as they choose and can continue to improve until they feel ready to submit it to their teacher. Furthermore, many teachers would prefer to focus on higher-level concerns, such as argumentation and voice when responding to student writing, and automated feedback has the potential to remove some of the burden of giving feedback on grammar to give teachers more time for these higher-level concerns.

Xi (2010) provides the following list of questions that can serve as a guide for evaluating automated feedback systems (see Table 3.2). Note that several of these questions could be asked, and have been asked, of feedback in general, as discussed above. Recent

Table 3.2 Validity Questions Related to Automated Feedback

Inference	*Validity Questions*
Evaluation: the scores are accurate representations of the performance	Does the automated feedback system accurately identify learner performance characteristics or errors?
Generalization: the observed score is an appropriate estimate of other scores obtained from other, similar observations	Does the automated scoring feedback system consistently identify learner performance characteristics or errors across performance samples?
Explanation: the scores can be attributed to the construct	Is the automated feedback meaningful to students' learning?
Utilization: the scores provide useful information for decisions and curriculum	Does the automated feedback lead to improvements in learners' performances? Does the automated feedback lead to gains in targeted areas of language ability that are sustainable in the long term? Does the automated feedback have a positive impact on teaching and learning?

Source: Xi, 2010.

research on automated feedback has focused on two questions: the accuracy of automated feedback, and its usefulness to students. I will discuss each of these questions briefly.

Does the Automated Feedback System Accurately Identify Learner Performance Characteristics or Errors?

The most common type of automated feedback is feedback on sentence-level errors (e.g., grammatical and mechanical errors), as opposed to content and organization. If automated feedback on errors is to be useful, it should be able to correctly identify the kinds of errors that ESL students are likely to make, and recent advances in AES systems have improved this ability, particularly in the areas of articles and preposition usage. Equally importantly, an automated feedback engine should be able to identify errors that are important for students to address, given the multiplicity of things that students need to think about when revising their writing. The general consensus is that the most important errors to address are those that (a) interfere with comprehension of the message; (b) are frequent; (c) are important for the particular genre or in the context of a particular teaching situation; and (d) are ones that the student is developmentally ready to address (Bates, Lane, & Lange, 1993; D. Ferris, 2011; Frodeson & Holten, 2003; Williams, 2005). While a computer can be programmed to identify errors that are likely to cause comprehension problems (for example, wrong words and verb formation errors tend to be more serious than article errors) and errors that occur frequently, the other considerations require teacher input and interpretation and may not be as easy to automate.

As noted above, automated scoring systems, like people, cannot always categorize ESL errors reliably, particularly in cases of garbled syntax, where the overall sense is clear but trained humans may not agree on the appropriate correction. For example, "he lead a good life" can be interpreted as a subject–verb agreement error or a tense error, depending on the context. Another problem is that humans do not always agree on whether or not something is an error. Preposition errors fall into this category (Chodorow et al., 2010; Tetreault & Chodorow, 2008) because of the variety of factors that influence their use and variability in their usage among native speakers. Given these complications, creators of AES algorithms need to decide whether false positives (identifying something as an error is actually not) or false negatives (failing to identify an actual error) are the greater problem. Chodorow et al. (2010) report that the developers of Criterion decided to minimize false positives; as a result, Criterion is able to identify with 80% accuracy only 25% of preposition errors. With article errors, these figures are slightly better: identifying 40% of errors with 90% accuracy. Recall that Ferris (2006) found that teachers agreed with trained raters 80–90% of the time, but this count included all errors, not only preposition and article errors, which are somewhat more difficult to categorize than errors such as verb tense or word form.

Thus, if e-rater can be considered the most advanced AES system for non-native writing, it appears that automated scoring engines have a similar accuracy rate as human raters, but may not identify as many errors as human raters do. Grimes and Warschauer (2010), commenting on low rates of identifying errors in MyAccess!, argue that automated feedback, even if imperfect, serves a useful pointing function, in that it identifies structures that are likely to be errors, thus reducing cognitive load on the writer.

It should also be noted that imperfect targeted feedback may have different consequences for learners at different levels of ability. In a study of Microsoft Word's spelling and grammar checker, Galleta, Durcikova, Everard and Jones (2005) found an interaction between verbal ability and use of software. In cases where the software correctly

identified errors, both high and low ability students performed equally well. In the case of false positives (errors falsely identified by the software) both groups of students were influenced by the incorrect error messages and wrongly corrected them. In the case of false negatives (errors not detected by the software) the high ability students' performance was particularly degraded; these students were much more likely to detect such errors when the software was turned off than when it was turned on. This result suggests that high ability students rely on the software too much and have an unrealistic expectation of how credible the software is. While this was only a single study, it serves as a reminder that teachers and students should not rely too heavily on automated feedback alone but must always interpret it.

Is the Automated Feedback Meaningful to Students' Learning?

Another critical question regarding automated feedback is whether it is effective in helping students improve their writing. Some recent research has begun to answer this question by investigating students' use of such feedback in their classrooms. Two studies in particular are relevant. Chen and Cheng (2008) studied the classroom use of My Access!, which provides both a score and specific feedback, in three university classrooms in Taiwan. The teachers received only one hour of training and were given free rein in determining how the software would be used. In the class where automated feedback was most well received by students according to a survey, the teacher used the software for formative assessment only, requiring that students achieve a certain computer score before turning their drafts in for teacher comments. In the second class, the teacher stopped using the software after six weeks of class, stating that the automated feedback was unhelpful and required more work on her part to help students interpret and use it, and also citing technical difficulties using the software. In the third class, the teacher used both the feedback and grading components of the software but did not provide much guidance to students on how to make use of the feedback features. Students in all classes had negative reactions towards the scoring features, and only some students found the feedback features useful. Students found in some cases that the feedback helped them attend to specific language problems such as sentence variety and the use of the passive voice, but they also felt the feedback was "vague," "abstract," and "repetitive" and did not help them revise their essays, particularly in development and coherence. Moreover, they felt that the automated system discouraged creativity in writing and overemphasized the use of transition words. They also thought that automated feedback would be more appropriate for students at lower levels of language proficiency who were still learning basic writing skills.

Grimes and Warschauer (2010) studied the implementation of MyAccess! in eight middle schools in Southern California. They found that the use of the software simplified classroom management and increased student motivation to write. However, some of their findings regarding automated feedback support the notion that teacher support is essential for effective implementation of automated feedback. They found that MyAccess! provided too much low-level feedback for students to absorb, and that they needed to supplement the feedback with handouts or checklists to help students prioritize and use the feedback. They also reported that some of the feedback (e.g., "adverb placement") was difficult for students to interpret, particularly ELLs.

Although the research in this area is limited, it does suggest several strategies for implementing automated feedback successfully. First of all, teacher training and support is crucial. Grimes and Warschauer (2010) report that lack of administrative support was one reason that automated scoring was discontinued in one district; similarly, the one hour

of training received by teachers in Chen and Cheng's (2008) study was not sufficient to make teachers feel confident in their use of the software. Teachers need to be trained in the technical aspects of using the system and sufficient support must be available to deal with inevitable technological breakdowns. Teachers also need to know how to make best use of the features in the software for their specific classroom situation. For example, the strategy of requiring students to achieve a certain score from the automated system before turning a paper in to the teacher appeared to be a useful motivating strategy.

Second, it is incumbent upon teachers to help their students interpret and prioritize feedback. Both studies found that the amount of low-level feedback was overwhelming, and some teachers reported that their workload increased because of the need to help students interpret the feedback.

Third, students are likely to take on the attitudes towards AES systems that their teachers adopt (Chen & Cheng, 2008), so teachers should be careful about how they present the tools to their students. Finally, teachers should remember that the most common AES systems are a commercial product (Chapelle & Chung, 2010) and their claims of success deserve critical scrutiny from scholars and teachers.

Just as teachers need to decide on their own strategies for error correction and feedback in writing, they need to be careful in how they present and use automated feedback in their own classrooms. Perhaps automated feedback is most useful as a "third voice" (Myers, 2003) in student–teacher conferences, where it provides some objective information about the essay and an evaluation based on that information with which the student and teacher can agree or disagree. As Grimes and Warschauer (2010, p. 34) state,

> Mindful use of [Automated Writing Evaluation] AWE can help motivate students to write and revise, increase writing practice, and allow teachers to focus on higher level concerns instead of writing mechanics. However, those benefits require sensible teachers who integrate AWE into a broader writing program emphasizing authentic communication, and who can help students recognize and compensate for the limitations of software that appears more intelligent at first than on deeper inspection.

While Grimes and Warschauer were addressing teachers of native speakers, their advice is equally valid for those who teach ELLs.

CONCLUSION

Despite calls by some to eliminate or reduce the use of AES systems in teaching and evaluating writing, it would be naïve to suggest that AES systems will not be used for scoring and feedback in the future. AES is here to stay, and the focus should be on continuing to improve both the human and technological sides of the equation. On the technological side, recent insights into differences between native and non-native writing and between lower- and higher-level proficiency writers that go beyond error counts may help improve automated scoring engines used for ELL writing. For example, Crossley and McNamara (2011) found significant differences between native and non-native college-level writers on four word-based indices (hypernymy, polysemy, lexical diversity, and stem overlap), suggesting that these indices may be useful indicators of native vs. ESL writing. Similarly, Friginal and Weigle (2012) found that co-occurrence of specific language features such as agentless passives, attributive adjectives, and lack of other features such as second person pronouns, mental verbs, and that-complement clauses were associated with higher essay

scores, suggesting that as students gain proficiency, their academic writing becomes more informational and less descriptive.

As for the human side of the equation, it has been shown that failures in the implementation of educational technology are often due to teacher resistance rather than problems with technology (Curan, 2003, 2005, cited in Grimes & Warschauer, 2010). Administrative and peer support, training, and willingness of teachers to experiment can be important in reducing such resistance. One possible path is to expose teachers to non-commercially produced automated tools that can be used to explore dimensions of their own students' writing. For example, Cotos (2012) reports improvement in writing among students using the Intelligent Academic Discourse Evaluator (IADE), a web-based, automated writing evaluation program that analyzes the discourse elements in the introduction section of research articles. Other corpus tools may be used by researchers or teachers wishing to do specific analyses on their own students' data. For example, the Gramulator (McCarthy, Watanabe, & Lamkin, 2012), can be used to identify the linguistic differences between two data sets—expert and novice papers, for example—and the LexTutor website (http://www.lextutor.ca/) includes numerous tools for analyzing vocabulary use automatically. Teachers who are familiar with such tools and can see their advantages and disadvantages may be more comfortable with commercially produced automated evaluation tools.

To quote Stewart Brand, editor of the *Whole Earth* journal: "Once a new technology rolls over you, if you're not part of the steamroller, you're part of the road" (http://famousquoteshomepage.com/2012/02/01/stewart-brand/). (Teachers of ELLs would do well to make sure they are part of the steamroller when it comes to automated essay evaluation.)

NOTE

1 ELL and NNS are used interchangeably in this chapter. Other terms frequently encountered in the literature are English as a second language (ESL) or Limited English Proficiency (LEP) students. For a useful overview of terminology, see Wolf and Farnsworth (in press).

REFERENCES

Attali, Y., & Burstein, J. (2006). Automated essay scoring With e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3), Available from http://www.jtla.org.

Bates, L., Lane, J., & Lange, E. (1993). *Writing clearly: Responding to ESL compositions*. Boston, MA: Heinle.

Burstein, J. (2002). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. *Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*.

California, Regents of the University of. (2012). University of California Office of the President Student Affairs: Entry Level Writing Requirement, from http://www.ucop.edu/elwr/index.html

Carson, J. (2001). A task analysis of reading and writing in academic contexts. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections* (pp. 246–270). Ann Arbor, MI: University of Michigan Press.

Chapelle, C., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Taylor & Francis.

Chapelle, C. A., & Chung, Y. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing, 27*(3), 301–315.

Chen, C.-F., & Cheng, W.-Y. (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.

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *English Journal, 93*, 47–52.

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing, 27*(3), 419–436.

Condon, W. (2006). Why less is not more: What we lose by letting the computer score writing samples. In P. F. Ericsson, & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 211–220). Logan, UT: Utah State University Press.

Condon, W. (2009). Looking beyond judging and ranking: Writing assessment as a generative practice. *Assessing Writing, 14*(3), 141–156.

Conference on College Composition and Communication. (2004). *CCCC position statement on teaching, learning, and assessing writing in digital environments*, February 25. Retrieved from http://www.ncte.org/cccc/resources/positions/digitalenvironments

Connor-Linton, J. (1993). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*(4), 762–765.

Cotos, E. (2012). Potential of automated writing evaluation feedback. *CALICO Journal, 28*(2), 420–459.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning, 21*(3), 170–191.

Crusan, D. (2010). *Assessment in the second language writing classroom.* Ann Arbor, MI: University of Michigan Press.

Cumming, A. (1989). Writing expertise and second-language proficiency. *Language Learning 39*(1), 81–141.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing 7*, 31–51.

Deane, P. (in press). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*.

Elliot, S. (2003). IntelliMetric: From here to validity. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahwah, NJ: Erlbaum Associates.

Ellis, R. (2005). Principles of instructed language learning. *System 33*(2), 209–224.

Emig, J. (1971). *The composing processes of twelfth graders*. Urbana, IL: National Council of Teachers of English.

Educational Testing Service. (2012). About the TOEFL iBT Test. Retrieved December 14, 2012, from http://www.ets.org/toefl/ibt/about/

Europe, Council of. (2001). *Common European framework for reference for languages: Learning, teaching assessment*. Cambridge: Cambridge University Press.

Feak, C., & Dobson, B. (1996). Building on the impromptu: A source-based writing assessment. *College ESL, 6*(1), 73–84.

Ferris, D. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81–104). Cambridge, UK: Cambridge University Press.

Ferris, D. (2011). *Treatment of error in second language student writing.* Ann Arbor, MI: University of Michigan Press.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 22*, 365–387.

Friginal, E., & Weigle, S. C. (2012). *Exploring multiple profiles of academic writing using corpus-*

based, multi-dimensional and cluster analysis. Paper presented at the Georgetown University Roundtable (GURT), Washington, DC.

Frodeson, J., & Holten, C. (2003). Grammar and the ESL writing class. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 141–161). Cambridge, UK: Cambridge University Press.

Galleta, D. F., Durcikova, A., Everard, A., & Jones, B. (2005). Does spell-checking software need a warning label? *Communications of the ACM, 48*(7), 82–85.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6). Retrieved September 1, 2012, from http://www.jtla.org

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63*(4), 480–499.

Homburg, T. J. (1984). Holistic evaluation of ESL composition: Can it be validated objectively? *TESOL Quarterly, 18*(1), 87–107.

Huot, B. (1993). The influence of holistic scoring preocedures on reading and rating student essays. In M. Williamson & Brian Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206–236). Cresskill, NJ: Hampton Press.

Kachru, B. B. (1997). World Englishes and English-using communities. *Annual Review of Applied Linguistics, 17*, 66–87.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527–535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement (4th ed)*. (pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kroll, B. (2001). Considerations for teaching an ESL/EFL writing course. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd ed.) (pp. 219–232). Boston, MA: Heinle.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*, 246–276.

Marsh, H.W. & Ireland, R. (1987). The assessment of writing effectiveness: A multidimensional perspective. *Australian Journal of Psychology, 39*, 353–367.

McCarthy, P., Watanabe, S., & Lamkin, T. (2012). *The gramulator: A tool to identify differential linguistic features of correlative text types*. Hershey, PA: IGI Global.

McKay, S. L. (2002). *Teaching English as an international language*. Oxford, UK: Oxford University Press

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication, 27*(1), 57–86.

Milanovic, M., Saville, N., & Shen, S. (1996). *A study of the decision-making behaviour of composition markers* Paper presented at the 15th Language Testing Research Colloquium, Cambridge and Arnhem.

Myers, M. (2003). What can computers contribute to a k-12 writing program? In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 3–20). Mahwah, NJ: Lawrence Erlbaum Associates.

No Child Left Behind Act, 115 Stat. 1425 (2001).

Nold, E. W., & Freedman, S. W. (1977). An analysis of readers' responses to essays. *Research in the Teaching of English, 11*, 164–174.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Roca de Larios, J. R., Murphy, L., & Marín, J. (2002). A critical examination of L2 writing process research. *Studies in Writing, 11*, 11–47.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly, 27*(4), 657–677.

Tetreault, J., & Chodorow, M. (2008). *The ups and downs of preposition error detection*. Paper presented at the COLING, Manchester, UK.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*, 327–369.

Truscott, J. (1999). The case for grammar correction in L2 writing classes: A response to Ferris. *Journal of Second Language Writing, 8*, 111–122.

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*, 4, 255–272.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex.

Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly, 45*, 769–774.

Warschauer, M., & Grimes, D. (2008). Automated writing in the classroom. *Pedagogies: An International Journal, 3*, 22–26.

Weigle, S. C. (2002). *Assessing writing*. New York: Cambridge University Press.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*(1), 27–55.

Weigle, S. C. (2005). Second language writing expertise. In K. Johnson (Ed.), *Expertise in language learning and teaching* (pp. 128–149). Hampshire, England: Palgrave Macmillan.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335–353.

Weigle, S. C. (2011). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. TOEFL iBT Research Report TOEFL iBT-15. Princeton, NJ: Educational Testing Service.

Weigle, S. C. & Montee, M. (in press). Raters' perceptions of textual borrowing in integrated writing tasks. In M. Tillema, E. Van Steendam, G. Rijlaarsdam. & H. van den Bergh (Eds.) *Measuring writing: Recent insights into theory, methodology and practices*. Bingley, UK: Emerald Books.

Williams, J. (2005). *Teaching writing in second and foreign language classrooms*. Boston, MA: McGraw-Hill.

Wolf, M., & Farnsworth, T. (in press). The use of English proficiency assessments for exiting English learners from ESL services: Issues and validity considerations. In A. Kunnan (Ed.), *The companion to language assessment*. Hoboken, NJ: Wiley-Blackwell.

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

You, X. (2004). The choice made from no choice: English writing instruction in a Chinese University. *Journal of Second Language Writing, 13*, 97–110.

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly, 19*, 79–102.

Zheng, Y., & Cheng, L. (2008). College English Test (CET) in China. *Lanuage Testing, 25*(3), 408–417.