CHAPTER 21

Computer-Based Writing Instruction

Laura K. Allen, Matthew E. Jacovina,
and Danielle S. McNamara

The development of strong writing skills is a critical (and somewhat obvious) goal within the classroom. Individuals across the world are now expected to reach a high level of writing proficiency to achieve success in both academic settings and the workplace (Geiser & Studley, 2001; Powell, 2009; Sharp, 2007). Unfortunately, strong writing skills are difficult to develop, as they require individuals to coordinate a number of cognitive skills and knowledge sources through the process of setting goals, solving problems, and strategically managing their memory resources (Flower & Hayes, 1980; Hayes, 1996). Given the difficulty of this process, students frequently underachieve on national assessments of writing proficiency (National Assessment of Educational Progress, 2007, 2011).

The successful acquisition of this complex set of skills largely depends on the instruction and feedback provided to students as they develop. Previous research suggests that writing proficiency is best enhanced through strategy instruction (Graham & Perin, 2007), along with extended practice and individualized feedback (Kellogg & Raulerson, 2007). Importantly, curricula aimed to meet these goals can be extremely difficult to implement. This method of writing instruction demands a significant amount of time from teachers, ranging from the time required to prepare materials and instructional content to the process of reading, editing, and providing individualized feedback on students' essays. Unfortunately, teachers rarely have enough time to devote to this process, as they are faced with increasingly large class sizes and, as a consequence, have reduced time for classroom instruction and planning (The National Commission on Writing, 2004).

In response to these issues surrounding effective writing pedagogy, there has been an increased effort to develop computer-based systems for writing assessment and instruction (Shermis & Burstein, 2003, 2013). These systems vary widely in their primary purposes, from the automated scoring of student essays to the provision of formative feedback or the explicit instruction of writing knowledge and strategies (Dikli, 2006; Graesser & McNamara, 2012; Roscoe, Allen, Weston, Crossley, & McNamara, 2014; Weigle, 2013; Xi, 2010). Despite the importance of writing strategy *instruction* and *feedback,* the majority of research conducted on this topic has focused on the development of computer-based systems that can provide reliable and valid scores to students' essays. However, more recently, researchers have placed a stronger emphasis on the development of computer-based systems that incorporate more instructional materi-

als, such as formative feedback and explicit instruction on the writing process (Roscoe et al., 2011). In this chapter, we provide an overview of computer-based tools and techniques that are currently being used to support writing instruction and practice. Additionally, we discuss more recent advances in this field with suggestions for future research and development.

Automated Essay Scoring

Automated essay scoring (AES) systems are the most prominent among computer-based writing tools. AES systems are technologies that allow computers to automatically evaluate the content, structure, and quality of written prose (Shermis & Barrera, 2002; Shermis & Burstein, 2003, 2013). In line with this goal, AES has been advertised as an efficient means through which large corpora can be rapidly scored, enabling writing instructors to assign more writing practice to their students without significantly adding to their workload (Dikli, 2006; Page, 2003; Shermis & Burstein, 2003, 2013). Further, large-scale testing services utilize AES systems to score writing tasks for high-stakes writing assessments, such as the Graduate Record Examination (GRE) or the Test of English as a Foreign Language (TOEFL; Dikli, 2006). Importantly, these systems do not provide formative feedback on students' essays, nor do they include instructional materials. Rather, their sole purpose is to serve as a valid and reliable alternative to human scoring that is both cost and time effective (Bereiter, 2003; Myers, 2003; Page, 2003).

AES Techniques

AES systems employ a diverse set of techniques to assign holistic grades to students' writing, including statistical modeling, Bayesian text classification, natural language processing (NLP), artificial intelligence (AI), and latent semantic analysis (LSA) (Deane, 2013; Dikli, 2006; Shermis & Burstein, 2003). In general, these methods all rely on the extraction of linguistic and semantic characteristics of a text to calculate essay scores. However, each AES system differs slightly in terms of the specific scoring methods and techniques employed. In particular, the most common techniques used by AES systems are those that leverage NLP techniques and those that utilize LSA.

e-rater (Burstein, 2003), IntelliMetric (Rudner, Garcia, & Welch, 2006), and the Writing Pal (W-Pal) (McNamara, Crossley, & Roscoe, 2013) are a few examples of computer-based writing systems that employ AES algorithms that rely on NLP techniques to score essays. NLP approaches to essay scoring are based on the assumption that essay quality can be evaluated using specific and measurable linguistic features (e.g., lexical diversity, average sentence length, referential cohesion) that are calculated using automated text analysis tools. By using a corpus of essays that have been scored by expert raters, a statistical model is built that combines linguistic features to create algorithms that assign scores to essays. These computational algorithms are tuned to match the expert raters' scores for the essays in the training corpus using statistical techniques such as machine learning algorithms, regression techniques, or Bayesian probabilities. The resulting algorithm is then applied to essays outside the training corpus (which then need not be scored by expert raters) to assign essay scores automatically.

The e-rater system, for example, uses 11 features of student essays (9 features related to writing properties and 2 related to content appropriateness) to assign holistic scores to student essays (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012). Each of these essay features encompasses at least one (but often more) underlying subfeature. For instance, the *style* feature contains subfeatures related to word repetition, as well as the use of inappropriate words and phrases. Within the e-rater system, each of these 11 features is assigned a weight that is determined either by its construct relevance or through the use of regression models that predict expert ratings of essays (Quinlan, Higgins, & Wolff, 2009). Holistic essay scores are then calculated using a weighted average of these feature scores (Enright & Quinlan, 2010).

Similarly, the AES engine within W-Pal assigns essay scores based on the calculation of numerous linguistic and semantic text features using Coh-Metrix (McNamara & Graesser, 2012; McNamara, Graesser,

McCarthy, & Cai, 2014) and the Writing Analysis Tool (WAT; McNamara et al., 2013). These indices are related to both lower-level aspects of student essays, such as the number of words or number of paragraphs in a text, and higher-level text features, such as semantic cohesion or the use of certain rhetorical devices. The algorithms implemented within W-Pal have been based on a number of different techniques (McNamara et al., 2013), and can be changed based on context or the age range of the students. One difference between W-Pal and other approaches such as e-rater is that W-Pal algorithms are built to generalize across topics or essay prompts. That is, the algorithms do not include features that are specific to an essay prompt, such as keywords, and the algorithms are constructed using a corpus of essays that vary in terms of their topic.

Another novel approach used within W-Pal is the use of hierarchical classification (McNamara, Crossley, Roscoe, Allen, & Dai, 2015). Accordingly, different linguistic features are combined to classify essays at different levels. At the first level, essays are divided into two groups of longer and shorter essays. The underlying assumption is that different linguistic features should predict the quality of longer versus shorter essays. The quality of the essays is then predicted by dividing the essays into subgroups or hierarchical levels. In the end, each essay score (e.g., 1–6) is based on a different set of algorithms that are iteratively applied, each using different linguistic features as well as different weights.

In contrast to NLP-based AES systems, the intelligent essay assessor (IEA; Landauer, Laham, & Foltz, 2003) utilizes LSA to score essays. LSA is a statistical technique that utilizes large corpora of documents to develop representations of world knowledge (Landauer, McNamara, Dennis, & Kintsch, 2007). Documents are represented in matrices, where each row represents a word and each column represents the context (i.e., the document) of that word. Accordingly, the individual cells represent the frequency of a word in a given context. A mathematical technique called singular value decomposition (SVD) is then used to reduce the number of columns in the matrix while maintaining the structure among the rows. Words can then be compared by calculating the cosine of the angle between the vectors formed by two rows. Cosine values approaching 1.0 represent high similarity between words, and values approaching 0 represent high dissimilarity.

The assumption of LSA-based scoring engines is that word meanings are determined by their co-occurrence with other words. Thus, it should be possible to extract semantic information from texts using word co-occurrence information from other texts. Similar to NLP-based scoring engines, IEA relies on a corpus of expert-scored essays to provide automated scores of new essays. The difference, however, is that the IEA system relies on *semantic similarity* among texts to assign essay scores, rather than linguistic text properties (Landauer et al., 2003; Rudner & Gagne, 2001; Streeter, Psotka, Laham, & MacCuish, 2004). Thus, essays are assigned high scores to the degree that they are semantically similar to other essays from the training set. Essays that are similar to higher-quality essays receive a high score, whereas essays that are more similar to lower-quality essays receive a lower score. One potential advantage of LSA-based scoring engines is that they do not require automated taggers and parsers, which are computationally heavy and difficult to implement. Further, these engines do not rely on weighted statistical models; instead, the original corpus serves as the model for essay quality. A disadvantage is that a corpus of scored essays is required for each essay prompt or essay topic, though this corpus can usually be much smaller than that required to construct a reliable statistical model using the linguistic features of the essays.

AES Reliability and Accuracy

Across these (and other) various techniques, AES systems tend to report fairly high reliability and accuracy (Attali, 2004; Landauer, Laham, Rehder, & Schreiner, 1997; Landauer et al., 2003; Warschauer & Ware, 2006). Specifically, expert human and computer scores tend to correlate between $r = .80$ and .85, which is similar to the range found between two human raters (Landauer et al., 2003; Rudner et al., 2006; Warschauer & Ware, 2006). For instance, Attali and Burstein (2006) assessed the reliability and true

score correlations between human scores and e-rater scores. They found that e-rater was *more* reliable than a single human and exhibited a true score correlation with a human rater at rho = .97.

In addition to correlation analyses, the reliability and accuracy of AES systems are also evaluated according to the percent of agreement between raters. Percent agreement is commonly reported in two forms: perfect agreement and adjacent agreement. Perfect agreement refers to an exact match between human and automated scores, whereas adjacent agreement refers to scores that are within 1 point of each other. Studies examining the accuracy of AES systems tend to report perfect agreement between 40 and 60% and adjacent agreement between 90 and 100% (Attali & Burstein, 2006; Dikli, 2006; McNamara et al., 2013; Rudner et al., 2006). Rudner and colleagues (2006), for instance, examined the accuracy of IntelliMetric in two separate studies and reported perfect agreement between 42 and 65% and adjacent agreement between 92 and 100%. Attali (2008) compared the agreement between two human raters to the agreement between e-rater and one human rater. He found that the two human raters had 56% perfect agreement and 97% adjacent agreement, whereas e-rater and one human rater had 57% perfect agreement and 98% had adjacent agreement. Overall, the results of these and other studies point to the strength of AES systems in their ability to provide reliable and accurate scores of essay quality.

Criticisms of AES

Despite the *accuracy* of many AES systems, the use of these systems on standardized assessments and in school classrooms has been met with a great deal of resistance (Condon, 2013; Deane, 2013). Some researchers have suggested that the systems do not assign scores with adequate accuracy, citing analyses that have shown systems to systematically over- or underestimate human ratings of essay quality (Wang & Brown, 2007). Further criticisms of AES systems have centered on students' ability to "game the system" by using their knowledge of the computerized scoring process. Powers, Burstein, Chodorow, Fowles, and Kukich (2002) explored this issue and found that students could cheat the e-rater system through the repetition of the same paragraph throughout the text, the use of complex sentence structures, and the inclusion of relevant content words. Thus, even if an essay was illogical, it could receive a high score if the linguistic features of the essay mirrored those in the algorithm.

Additional criticisms have centered on the content of the essays written in the AES systems. Importantly, not all genres of writing will be scored accurately using the same AES algorithms. For instance, essays that require students to integrate multiple assigned documents have specific scoring considerations that are different from argumentative SAT-style essays. Britt, Wiemer-Hastings, Larson, and Perfetti (2004) used several techniques to identify problems in students' source-based essays. As an example, they used LSA to determine if sentences from students' papers overlapped with any sentences from the assigned documents. If the LSA cosine threshold was not met for at least three different sources, students' papers were flagged as not integrating an appropriate number of sources—and those students would subsequently receive feedback on how to integrate sources. Although this approach presents a strong foundation on which to begin to study source-based writing, it is evident that the automated evaluation of certain writing genres may be more complex and thus require a great deal more research and analysis to be considered valid and reliable.

Perhaps the most significant criticism met by AES systems pertains to the validity of the essay quality assessments (Cheville, 2004, Clauser, Kane, & Swanson, 2002; Condon, 2013; Deane, 2013; Ericsson & Haswell, 2006). As previously mentioned, many AES systems rely heavily on the use of linguistic features to assess essay quality (McNamara, Crossley, & McCarthy, 2010). This list of potential features is expansive—ranging from lower-level text information, such as sentence length and word frequency, to high-level features, such as rhetorical strategies and semantic cohesion (Landauer et al., 2007; McNamara et al., 2014; Shermis & Burstein, 2003; Tausczik & Pennebaker, 2010). Using these tools, researchers have learned a great deal about the linguistic features that relate to higher essay qual-

ity, such as elaboration, organization, and lexical sophistication (Deane, 2013). However, AES systems do not currently have the capability of detecting more subtle and subjective features of students' essays. For instance, what can linguistic essay features tell us about students' creativity or the depth and development of their ideas and arguments? This and similar questions remain to be answered. In general, critics of AES tend to emphasize the point that these automated systems fail to measure *meaningful* aspects of text; rather, they only measure an extremely restricted portion of writing proficiency (Deane, 2013).

Automated Writing Evaluation

In contrast to AES systems, automated writing evaluation (AWE) systems provide students with feedback on their writing (Crossley, Varner, Roscoe, & McNamara, 2013; Grimes & Warschauer, 2010). The two major benefits of AWE systems is that they provide opportunities for students to practice writing, as well as to receive summative and formative feedback on their essays—all without the input of an instructor. AES elements of AWE systems provide the automated *scoring* of students' essays; however, AWE systems extend beyond this assessment purpose by providing writing instruction and feedback to students, as well as features for teachers that can aid in classroom management (e.g., class rosters or grade books). A number of these AWE systems have now been developed for use in writing classrooms, such as Criterion (scored by the e-rater AES system), MyAccess (scored by IntelliMetric), WriteToLearn (scored by Intelligent Essay Assessor), and WPP Online (scored by PEG).

Deliberate Writing Practice

One of the primary benefits of AES and AWE systems is that they provide more opportunities for students to practice writing. Engaging in extended deliberate practice (Ericsson, Krampe, & Tesch-Römer, 1993) is considered necessary for the development of successful writing skills (Johnstone, Ashbaugh, & Warfield, 2002; Kellogg & Raulerson, 2007), and, according to Kellogg and Raulerson (2007)'s review of the literature, this development takes approximately 10 years of practice. Further, Kellogg and Raulerson suggest that the *type* of writing practice students undertake is critical for their success. They warn against requiring "marathon" writing sessions. Instead, they suggest that writing practice must be *deliberate* in that students develop clear goals and receive formative feedback on their writing.

Unfortunately, given increasing class sizes, it is unreasonable to expect teachers to provide useful feedback to each student across multiple writing assignments. Given continued advancements in AWE scoring and feedback, computer-based writing systems may be able to act as supplements to traditional classroom instruction (Roscoe et al., 2014). Specifically in the case of AWE systems, students are afforded the opportunity to practice holistic essay writing. In these systems, students can complete as many writing assignments as they choose and, for each, receive feedback specific to the problems present in their essays. Further, students can take this feedback and revise their essays based on the suggestions provided by the system. This revision process allows students to engage in the iterative writing process without having to wait an extended period of time for input from the teacher (Roscoe et al., 2011; Roscoe, Varner, Crossley, & McNamara, 2013; Shute, 2008).

In addition to providing immediate scoring and feedback during practice, computer systems can help to promote students' persistence. Grimes and Warschauer (2010), for example, examined the use of the AWE MyAccess in middle schools over a 3-year period. Teachers reported that students using the system wrote more often, demonstrated greater autonomy in their writing development, and were more motivated to engage in writing practice. Among their explanations for these benefits, Grimes and Warschauer indicated that submitting early essay drafts to MyAccess had a lower risk than submitting an essay to the teacher. Instead of being judged by human graders—perhaps especially by the teacher who would be assigning the grade to the final product—the computer offered students helpful, yet unthreatening, feedback.

Formative Feedback

One of the most crucial components of computer-based writing instruction is the ability of the computer to provide accurate and formative feedback on students' writing. Because of the substantial effort required from instructors to provide writing feedback and guide students to effective practice activities based on that feedback, AWE systems can be attractive educational tools. One important question, however, regards the degree to which computer-based writing assessments enhance students' writing quality in the context of classroom instruction. A recent meta-analysis by Graham, Hebert, and Harris (in press) investigated the benefits of formative writing assessments directly tied to classroom instruction. Their analysis indicated that providing automated feedback to students significantly enhanced the quality of their writing, yielding an average-weighted effect size of 0.38. These findings support the notion that computer-based writing systems have the potential to improve students' writing, particularly if they provide formative feedback that is directly tied to instruction.

An inherent advantage of computer-based systems is the speed with which feedback can be delivered to students. Providing immediate formative feedback in a computer-based system can help students identify their strengths and weaknesses, keep them engaged in the learning process, and improve their learning outcomes (Gikandi, Morrow, & Davis, 2011). By contrast, students often receive feedback only on finished work in the context of more traditional classroom instruction. When feedback is delivered after students have moved on to a new assignment or topic, they are less likely to devote attention to understanding how that feedback could usefully improve their future writing (Frey & Fisher, 2013). Despite the clear advantage of immediate feedback, however, several challenges face computer-based writing systems in providing students with beneficial writing feedback (Roscoe et al., 2011; Shute, 2008). Such systems must present appropriate types of feedback, and must also offer methods that can address student weaknesses. AWE systems must also be designed such that feedback messages clearly relate to a student's work. Systems that provide generic feedback messages in response to students' essays are less effective at guiding students' revision process and teaching skills that can be applied during future writing (Chen & Cheng, 2008).

Despite the *importance* of feedback in computer-based writing systems, little research has been conducted to examine the usability and most effective forms of automated feedback (Roscoe et al., 2011; Roscoe et al., 2014). Traditional computer-based writing software often provides low-level feedback by focusing on the mechanical and grammatical errors in student essays. Although this type of feedback may improve the readability of students' writing, it does little to improve their overall writing skills (Crossley, Kyle, Allen, & McNamara, 2014; Graham & Perin, 2007). In line with critiques concerning the validity of meaningfulness of automated essay scores, the provision of higher-level feedback can provide strategies and techniques for writing that should ultimately prove more useful than lower-level feedback, particularly for developing writers. To address concerns about the potential usefulness of feedback, the W-Pal system provides students with high-level feedback that focuses on actionable strategies that students can use to revise their essays (Roscoe, Varner, et al., 2013). For example, if W-Pal's feedback algorithm classifies a student's argumentative essay as being poorly structured, it might provide suggestions about how to organize an essay using flowcharts that can visualize the structure of an essay.

Intelligent Tutoring Systems for Writing

Intelligent tutoring systems (ITSs) currently provide the most sophisticated form of computer-based writing instruction. ITSs are computer-based programs that have been designed to provide individualized instruction and feedback to students based on their needs. In well-defined domains such as mathematics and physics, ITSs have had success in modeling what students need to know and what they (seem to) actually know, and in providing specific problem sets and feedback

that adapt to their needs based on their performance (Beal, Arroyo, Cohen, & Woolf, 2010; Graesser et al., 2004). In fact, ITSs have had similar success in improving learning outcomes as human tutors (VanLehn, 2011). The architecture of a complete ITS includes an expert model, a student model, and tutorial strategies (Neuwirth, 1989). However, no system is perfectly adaptive, and even the most efficacious systems are continuously working to improve the implementation of these three components. For an ill-defined domain, such as writing, the challenge to provide personalized and adaptive instruction and feedback becomes even greater. Creating an expert model for how to compose a well-reasoned argumentative essay, for instance, is more complex than creating an expert model for how to solve a system of equations. Likewise, determining where a student is failing is more difficult for composition than mathematical problem solving.

Our discussion of the current state of AWE systems has already forecasted the major advancements in ITSs designed to improve students' writing, as well as many of the weaknesses that future work will attempt to address. Algorithms are built using expertly graded writing samples that can allow the accurate scoring of student writing. Additionally, these algorithms can identify the strengths and weaknesses of a given student's work. Together, these capabilities can be used by ITSs to build expert and student models.

Whereas AWE software is often presented as standalone software providing the opportunity to write essays and receive feedback, an ITS can offer a suite of instructional and practice lessons, with the AWE embedded within the system. ITS software can respond to several challenges in the writing domain, namely, the need for increased use of strategy instruction and strategy practice. Because a primary goal of educators is to provide formative assessment, a successful ITS for writing must be able to provide students with information that can guide their future composition. Thus, beyond what an AWE offers, an ITS for writing aims to deliver a more complete tutorial experience, providing students with writing strategies and goals (Roscoe & McNamara, 2013).

In the following sections, we describe how an ITS for writing can provide strategy instruction, promote extended practice, provide higher-level feedback, and individualize instruction to each student. This is not a comprehensive list of the desirable features of an ITS for writing. However, they represent a synergistic set of features that highlight how an ITS can provide a valuable educational package for adolescent writers.

Strategy Instruction

A crucial component of writing instruction is teaching strategies to students. In meta-analyses, strategy instruction is consistently shown to be one of the most effective means of improving adolescent writing (Graham, 2006; Graham & Perin, 2007). The strategy instruction included in these meta-analyses focuses on teaching explicit strategies for planning, revising, and/or editing an essay. Fidalgo, Torrance, and García (2008), for example, developed an intervention called cognitive self-regulation instruction (CSRI) that taught strategies for planning and revising. Two years after completing the CSRI, students' writing products and writing process differed from that of control students who did not receive the intervention. Students who received the training produced higher-quality texts that were better structured, and they reported spending more time outlining their writing during planning. Moreover, CSRI students were less likely to report a lack of motivation, and seemed to have higher writing self-efficacy, making fewer negative comments about their writing. Overall, writing strategy instructions appears to support long-lasting benefits to writing, influencing not only the overall quality of students' essays, but also their writing process (Braaksma, Rijlaarsdam, van den Bergh, & van Hout-Wolters, 2004; Torrance, Fidalgo, & García, 2007) and motivation to write (Graham, Harris, & Mason, 2005).

ITSs that intend to act as an effective learning tool for the entire writing process should therefore include strategy instruction. One approach to this goal is to use pedagogical agents to deliver instructional lessons that explain writing strategies and provide examples of how the strategies can

be used while planning, writing, or revising (Dai, Raine, Roscoe, Cai, & McNamara, 2011). In W-Pal, for example, animated agents explain several writing strategies throughout a series of eight modules corresponding to prewriting, writing, and revising. The conclusion building module presents the *RECAP* strategy, which advises writers to *r*estate their thesis, *e*xplain how their thesis was supported, *c*lose the essay, *a*void new arguments, and *p*resent their ideas in an interesting way. For each piece of advice, an animated agent explains the strategy's purpose and meaning and provides an example of how it can be implemented. For example, the lesson on restating the essay's thesis suggests paraphrasing strategies to change particular thesis statements from the introduction to fit into the conclusion.

Providing computer-based strategy instruction on its own is likely to have a positive influence on students' writing, but an ITS affords instructional designers additional opportunities to increase its effectiveness. For example, students benefit when they are aware that learning the presented strategies is important and when they receive feedback about how well they have learned the strategies (Graham & Perin, 2007). An ITS is able to provide tests of strategy acquisition that emphasize strategies and provide performance feedback, in addition to the holistic writing practice provided by traditional AWE systems.

Strategy instruction also provides a context for which formative feedback can be delivered in a more meaningful way. To be effective, formative feedback should relate to information that students are learning (Graham et al., in press), and ITSs can serve as an environment where this instruction and feedback can be integrated. W-Pal, for instance, provides essay feedback directly based on lesson videos and practice games. Students who are unsure about how to implement suggested strategies into their revisions and future writing are able to reengage with these materials. Students who receive feedback that their essay is unstructured might watch a lesson video on creating outlines and flowcharts, or play a practice game in which they unscramble pieces of someone else's outline, learning how to identify and organize important pieces of evidence.

Yet, even an ITS that provides timely, appropriate feedback that is supplemented by content within the system can be ineffective for certain learners. Some students are simply going to ignore feedback messages and will fail to adapt their writing (Wingate, 2010). But an ITS should not, in turn, ignore these students. Instead, a successful system should identify profiles of behavior and adjust instruction accordingly. Although no current ITS for writing instruction is able to do this satisfactorily, we discuss future avenues for this research later in this chapter. In particular, we define the goals for writing-based ITSs and describe how researchers might begin to approach such an overwhelming, yet clearly important, objective.

Specialized Modes of Writing Practice

A key component of ITSs is their ability to provide multiple forms of practice. Unlike AWE systems, which only provide holistic essay practice, ITSs can provide writing practice along with component-based practice to increase students' writing proficiency. In W-Pal, for example, students have the opportunity to engage in holistic essay practice, where they write entire SAT-style persuasive essays. Additionally, they are able to practice specific strategies they have learned in lesson videos in strategy-specific practice sections (Allen, Crossley, Snow, & McNamara, 2014; Roscoe & McNamara, 2013). After viewing instructional videos on conclusion building, students can engage in practice that requires them to identify problematic implementations of conclusion writing strategies, and to write conclusions to essays. In both cases, students receive feedback and are referred back to the lesson videos for additional help.

In addition to offering practice at multiple levels of specificity, ITSs can also add certain elements to increase motivation and persistence among students. Because students often become bored by extended practice in traditional ITSs, developers have begun to create game-based learning environments that leverage students' enjoyment of gaming (Jackson & McNamara, 2013). W-Pal uses game-based practice to encourage students' understanding of the writing strategies taught in the system (Allen, Crossley et al., 2014;

Roscoe, Brandon, Snow, & McNamara, 2013). For example, *Speech Writer* requires students to help a friend rewrite a speech he is giving for the debate team. While doing so, students indicate which strategies they are using to fix the speech. Students then receive points based on how well they implemented the strategies while editing the speech. The simple narrative provides a context for writing practice, and the points system provides feedback and can motivate students to play again to improve their score. Findings from high school students demonstrated the efficacy of engaging with the entire W-Pal system, including game-based strategy practice such as Speech Writer, compared to engaging solely in holistic essay writing (Allen, Crossley, et al., 2014; see also Roscoe & McNamara, 2013). Game-based strategy practice offers a concrete motivation for students to understand and remember strategies, while providing frequent feedback and performance measures through game scores and achievements.

Mounting evidence across educational domains suggests that well-designed game-based practice can be effective at increasing students' learning outcomes (Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013). A successful educational game can promote a "game cycle" in which players interact with the game (e.g., editing a text in Speech Writer), receive feedback, and are motivated to reengage based on their favorable judgments of the game (Garris, Ahlers, & Driskell, 2002). Games that require students to write and use writing strategies can be used to help meet Kellogg and Raulerson's (2007) goal of extended writing practice. Thus, when educational games are combined with more traditional writing practice within an ITS, students are able to engage in a variety of practice modes to improve their understanding of writing strategies and their composition skills. Of course, for practice to be optimally effective, students must receive timely, accurate, and appropriate feedback and have resources available to remedy their weaknesses. Notably, however, determining the appropriateness of feedback for an individual student is a considerable challenge. Next, we discuss how an ITS might individualize students' experience within a system to provide more effective writing instruction.

Individualizing Instruction

According to VanLehn (2006), to be considered a true ITS, a system should offer access to content and assistance in a way that is suitable for each student's knowledge state and the system's task domain. Within a given task, the system should behave appropriately in terms of *what type* of feedback and assistance should be given to a student, *when* it should be given, and *how* it should be given. Above, we described how feedback might be tailored and delivered to individual students based on the content of their writing, but we posed the problem of students who disregard that feedback. This problem can be decomposed into two issues: how to identify students who are not benefiting from feedback, and how to present learning materials and feedback in an alternative format.

Although the complex nature of assessing writing compared to other domains (e.g., math) is often described as a challenge, writing samples also provides an abundance of information about students that can help guide the adaptive behavior of an ITS. As NLP algorithms are refined and become more successful at identifying weaknesses in students' writing, comparisons between drafts of an essay and its revision, and even between an essay and subsequent writing, can begin to assess the success students have in following suggestions from feedback messages. If students are not benefiting from the default method of feedback delivery, the system should try something else. Existing ITSs already have multiple formats of instruction, including lesson videos and practice activities; if a student fails to respond to written feedback messages but scores highly on practice games, the system might assign an appropriate strategy game instead.

Much additional research needs to be conducted before seemingly straightforward methods of individualizing writing instruction and feedback (such as above) can be effectively implemented. For example, forcing activities on students is likely to reduce their perceived control over the system, which could lead to fewer positive emotional responses (Pekrun, 2006). Therefore, researchers must first identify system designs that can subtly encourage students to complete certain activities without overtly

removing students' ability to control the system. Additional research must focus on how to analyze students' performance and behaviors within systems as a means to assess their learning styles and instructional needs in less intrusive ways (e.g., Snow, Likens, Jackson, & McNamara, 2013). Through such work, systems will be able to obtain vital information about students without disrupting instruction by asking students to complete surveys designed to capture individual differences such as motivation or cognitive flexibility.

To illustrate, a recent study examined 16 essays written by individual students in the W-Pal system, covering several different topics (Allen, Snow, & McNamara, 2014). An analysis of the degree of cohesion in each essay revealed that students were more or less flexible in their use of cohesion across different essays. More skilled writers demonstrated greater flexibility in their use of cohesion, whereas less skilled writers employed cohesion more rigidly across the wide array of essay topics. These results can inform the individualization of instruction. Specifically, when students do not vary their writing, they may benefit from instruction to change their approach based on the given prompt. More generally, analyses of students' writing patterns can be useful in determining appropriate essay feedback. Through analysis of students' behaviors and performance while using a system, profiles for each student can be built nonintrusively, and instruction can be more successfully individualized.

Conclusion

In this chapter, we have described a wealth of research that has been conducted to develop and test computer systems for writing instruction. In terms of small and large-scale assessments, computers can help teachers and testing services by providing valid and reliable ratings and feedback on students' essays. These automated systems can then provide students with significantly more opportunities to practice their writing along with suggestions and strategies for how to revise their essays and develop their writing skills. More recently, researchers and educators have moved toward using computers as methods for providing adaptive and personalized writing instruction. Intelligent tutoring systems can provide context for the scores and feedback that students receive on their essays and allow them to receive explicit instruction and practice in areas where they need the most help.

Despite these advancements in the field, many questions remain unanswered. For instance, can computers provide valid writing assessments when the *content* of the essays is the principal component of the essay (e.g., science reports or history papers)? Similarly, is computer-based language assessment limited to argumentative styles of writing, or can more subjective and creative forms of writing be similarly measured? These questions and many more remain to be explored in the future. As technology improves and as more research accumulates, we can begin to move toward finding answers to these questions and developing more sophisticated tools to support the successful development of students' writing skills.

Acknowledgments

This work was supported by the Institute of Education Sciences (IES), United States Department of Education, through Grant R305A120707 to Arizona State University. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily represent the views of the IES.

References

Allen, L. K., Crossley, S. A., Snow, E. L., & McNamara, D. S. (2014). L2 writing practice: Game enjoyment as a key to engagement. *Language Learning and Technology, 18*, 124–150.

Allen, L. K., Snow, E. L., & McNamara, D. S. (2014). The long and winding road: Investigating the differential writing patterns of high and low skilled writers. In J. Stamper, S. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 304–307), London, UK.

Attali, Y. (2004, April). *Exploring the feedback and revision features of Criterion*. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.

Attali, Y. (2008). *E-rater performance for*

TOEFL iBT independent essays. Unpublished manuscript.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2.0. *Journal of Technology, Learning and Assessment, 4* (np).

Beal, C., Arroyo, I., Cohen, P., & Woolf, B. (2010). Evaluation of AnimalWatch: An intelligent tutoring system for arithmetic and fractions. *Journal of Interactive Online Learning, 9*, 64–77.

Bereiter, C. (2003). Automated essay scoring's coming-of-age. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary approach.* Mahwah, NJ: Erlbaum.

Braaksma, M. A. H., Rijlaarsdam, G., van den Bergh, H., & van Hout-Wolters, B. H. A. M. (2004). Observational learning and its effects on the orchestration of writing processes. *Cognition and Instruction, 22*, 1–36.

Britt, A., Wiemer-Hastings, P., Larson, A., & Perfetti, C. A. (2004). Using intelligent feedback to improve sourcing and integration in students' essays. *International Journal of Artificial Intelligence in Education, 14*, 359–374.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary approach* (pp. 113–121). Mahwah, NJ: Erlbaum.

Chen, C. F. E., & Cheng, W. Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning and Technology, 12*, 94–112.

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

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education, 15*, 413–432.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*, 100–108.

Crossley, S. A., Kyle, K., Allen, L. K., & McNamara, D. S. (2014). The importance of grammar and mechanics in writing assessment and instruction: Evidence from data mining. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 300–303), London, UK.

Crossley, S. A., Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Using automated cohesion indices as a measure of writing growth in intelligent tutoring systems and automated essay writing systems. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED)* (pp. 269–278). Heidelberg, Germany: Springer.

Dai, J., Raine, R. B., Roscoe, R., Cai, Z., & McNamara, D. S. (2011). The Writing-Pal tutoring system: Development and design. *Journal of Engineering and Computer Innovations, 2*, 1–11.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7–24.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*, 3–35.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater(R) scoring. *Language Testing, 27*, 317–334.

Ericsson, P. F., & Haswell, R. (Eds.). (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Ericsson, K. A., Krampe, R. Th., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363–406.

Fidalgo, R., Torrance, M., & García, J. N. (2008). The long-term effects of strategy-focused writing instruction for grade six students. *Contemporary Educational Psychology, 33*, 672–693.

Flower, L., & Hayes, J. (1980). Identifying the organization of writing processes. In L. Gregg & E. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Hillsdale, NJ: Erlbaum.

Frey, N., & Fisher, D. (2013). A formative assessment system for writing improvement. *English Journal, 103*, 66–71.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming, 33*, 441–467.

Geiser, S., & Studley, R. (2001). *UC and SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California*. Oakland: University of California.

Gikandi, J. W., Morrow, D., & Davis, N. E.

(2011). Online formative assessment in higher education: A review of the literature. *Computers & Education, 57*, 2333–2351.

Graesser, A., Lu, S., Jackson, G., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments & Computers, 36*, 180–192.

Graesser, A. C., & McNamara, D. S. (2012). Reading instruction: Technology-based supports for classroom instruction. In C. Dede & J. Richards (Eds.), *Digital teaching platforms: Customizing classroom learning for each student* (pp. 71–87). New York: Teachers College Press.

Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187–207). New York: Guilford Press.

Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology, 30*, 207–241.

Graham, S., Hebert, M., & Harris, K. R. (in press). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*, 445–476.

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*. Retrieved January 5, 2012, from *www.jta.org*.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & L. S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 1–27). Hillsdale, NJ: Erlbaum.

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology, 105*, 1036–1049.

Johnstone, K. M., Ashbaugh, H., & Warfield, T. D. (2002). Effects of repeated practice and contextual-writing experiences on college students' writing skills. *Journal of Educational Psychology, 94*, 305–315.

Kellogg, R. T., & Raulerson, B. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review, 14*, 237–242.

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

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (pp. 412–417). Mahwah, NJ: Erlbaum.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

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

McNamara, D. S., Crossley, S. A., & Roscoe, R. D. (2013). Natural language processing in an intelligent writing strategy tutoring systems. *Behavior Research Methods, 45*, 499–515.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35–59. McNamara, D. S., & Graesser, A. C. (2012). CohMetrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188–205). Hershey, PA: IGI Global.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.

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 approach* (pp. 3–20). Mahwah, NJ: Erlbaum.

National Assessment of Educational Progress. (2007). The Nation's Report Card: Writing 2007. Retrieved November 5, 2010, from *nces.ed.gov/nationsreportcard/writing*.

National Assessment of Educational Progress. (2011). The Nation's Report Card: Writing 2011. Retrieved January 5, 2012, from *nces.ed.gov/nationsreportcard/writing*.

The National Commission on Writing. (2004). *Writing: A ticket to work. Or a ticket out: A survey of business leaders*. Retrieved from *www.collegeboard.com/prod_downloads/writingcom/writing-ticket-to-work.pdf*

Neuwirth, C. (1989). Intelligent tutoring systems: Exploring issues in learning and teaching writing. *Computers and the Humanities, 23*, 45–57.

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: Erlbaum.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*, 315–341.

Powell, P. (2009). Retention and writing instruction: Implications for access and pedagogy. *College Composition and Communication, 60*, 664–682.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18*, 103–134.

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct coverage of the e-rater® scoring engine* (ETS Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service.

Ramineni, C., Trapani, C. S., Williamson, D. M. W., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater® scoring engine for the TOEFL® independent and integrated prompts*. (ETS Research Report No. RR-12-06). Princeton, NJ: Educational Testing Service.

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition, 34*, 39–59.

Roscoe, R. D., Brandon, R. D., Snow, E. L., & McNamara, D. S. (2013). Game-based writing strategy practice with the Writing Pal. In K. Pytash & R. Ferdig (Eds.), *Exploring Technology for Writing and Writing Instruction* (pp. 1–20). Hershey, PA: IGI Global.

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology, 105*, 1010–1025.

Roscoe, R. D., Varner, L. K., Cai, Z., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2011). Internal usability testing of automated essay feedback in an intelligent writing tutor. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 543–548). Menlo Park, CA: AAAI Press.

Roscoe, R. D., Varner, L. K., Crossley, S. A., McNamara, D. S. (2013). Developing pedagogically-guided algorithms for intelligent writing feedback. *International Journal of Learning Technology, 8*, 362–381.

Rudner, L., & Gagne, P. (2001). *An overview of three approaches to scoring written essays by computer* (ERIC Digest number ED 458 290).

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric™ essay scoring system. *Journal of Technology, Learning, and Assessment, 4*(4).

Sharp, D. B. (2007). Learn to write. ISA Careers website. Retrieved September 20, 2011, from *www.isa.org/Template.cfm?Section=Careers&Template=/ContentManagement/ContentDisplay.cfm&ContentID=5328*.

Shermis, M. D., & Barrera, F. D. (2002). Automated essay scoring for electronic portfolios. *Assessment Update, 14*, 1–5.

Shermis, M., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

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

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189.

Snow, E. L., Likens, A., Jackson, G. T., & McNamara, D. S. (2013). Students' walk through tutoring: Using a random walk analysis to profile students. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 276–279). Heidelberg, Germany: Springer.

Streeter, L., Psotka, J., Laham, D., & MacCuish, D. (2004). *The credible grading machine: Essay scoring in the DOD* [Department of Defense]. The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), Orlando, FL.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and

computerized text analysis methods. *Journal of Language and Social Psychology, 29*, 24–54.

Torrance, M., Fidalgo, R., & García, J. N. (2007). The teachability and effectiveness of cognitive self-regulation in sixth-grade writers. *Learning and Instruction, 17*, 265–285.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*, 227–265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*, 197–221.

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment, 6*. Retrieved September 6, 2013, from *www.jtla.org*.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*, 157–180.

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 1*, 85–99.

Wingate, U. (2010). The impact of formative feedback on the development of academic writing. *Assessment & Evaluation in Higher Education, 35*, 519–533.

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*, 249–265.

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