Automated Essay Scoring and The Search for Valid Writing Assessment

Andrew Klobucar

New Jersey Institute of Technology, Newark, NJ

Norbert Elliot

New Jersey Institute of Technology, Newark, NJ

Paul Deane

Educational Testing Service, Princeton, NJ

Chaitanya Ramineni

Educational Testing Service, Princeton, NJ

Perry Deess

New Jersey Institute of Technology, Newark, NJ

Alex Rudniy

New Jersey Institute of Technology, Newark, NJ

Abstract

In this chapter, we present selected findings from a collaborative program of research between a non-profit educational research institution and a public science and technology university. After providing a background to the design of automated essay scoring (AES), we explore this method as one of six systems designed to assess writing performance. Because writing assessment is best validated in specific settings, we focus on a study undertaken in the fall of 2010 (N=1006) with the performance of first year admitted students. This case reveals that the best predictor of course grades is a model of measures embedded early in a semester that combine two AES timed writing submissions and a writing sample allowing 48-hour completion ($R = .475$, $R^2 = .225$, $p < 0.01$). End-of-semester portfolio assessment systems reveal the importance embedded measures providing substantial construct coverage. As discontinuous as AES is in a print environment, our framework demonstrates that this technology is congruent in a digital environment. The chapter concludes with a depiction of AES as a technology at a nascent stage of development. The future roles that writing assessment systems play depend on their integration into the practices of teachers and students.

Keywords: automated essay scoring, writing assessment, writing instruction

Representation, Automation, and Application: The Search for Valid Writing Assessment

In this chapter, colleagues at the Educational Testing Service (ETS) and New Jersey Institute of Technology (NJIT) present recent research aimed at unifying the instruction and assessment of writing. The research we present attends to the complexities involved in systems of writing assessment—from timed writing samples scored by machines to portfolios scored by humans. In the period from 2009 to the present writing, we have come to view engagement with automated essay scoring (AES) as permitting opportunities, as well as challenges, for instructors and their students.

Design of Automated Essay Scoring

Writing assessments often use constructed-response tasks. As Baldwin, Fowles, and Livingston (2005) have defined them, these tasks ask the examinee to display defined skills and knowledge. The performance, or response, may take many forms, from essays word-processed on a computer to the production of a course portfolio. Automated scoring systems have been developed for a variety of constructed-response item types, including essays (Shermis & Burstein, 2003), mathematical equations (Singley & Bennett, 1998), short written responses with well-defined correct answers (Leacock & Chodorow, 2003), and spoken responses (Xi, Higgins, Zechner, & Williamson, 2008). Assessment of the essay has attracted the most attention. More than 12 different automated essay evaluation systems have been developed, including Project Essay Grade (Page, 1966; 1968; 2003), engine 5 from Knowledge Analysis Technologies™ (Landauer, Laham, & Foltz, 2003), IntelliMetric™ (Rudner, Garcia, & Welch, 2006), and e-rater® (Attali & Burstein, 2006; Burstein, 2003). Each engine predicts human scores by modeling features of the written text and combining them using some statistical method.

Automated scoring can reproduce many of the advantages of multiple-choice scoring, including speed, consistency, transparent scoring logic, constant availability, lower per-unit costs, and the potential to provide detailed performance-specific feedback not feasible for human scoring under operational conditions.

In many contexts, use of automated methods to score writing is contested. Concerns range from the message AES use sends about the general nature of composition studies to the specific impact of the technology on writing instruction and student learning. The research reported here is not intended to address such controversies; rather, our focus is to explore ways in which automated essay scoring might fit within a larger ecology as one among a family of assessment techniques supporting the development of digitally enhanced literacy in its many forms. Viewed in this way, our work is responsive to a change in the nature of communication that is taking place within contemporary culture and which is certain to have profound ramifications for writing in academic environments.

The focus of the present collaboration is the Criterion Online Writing Evaluation Service (Burstein, Chodorow, & Leacock, 2003; Attali, 2004), an integrated assessment and instructional system that collects writing samples and provides instant scores and annotated feedback supported by an AES system, e-rater (Attali & Burstein, 2006). E-rater is a computer program that employs natural language processing (NLP) technology to extract defined features of writing and combine them in a regression model to score essays. Based in the results of e-rater, the Criterion platform then provides feedback that focuses on the defined features: grammar, usage and mechanics; style; and elements of essay structure. Based on continuous NLP development efforts, the set of features is periodically enhanced and updated.

Use of automated scoring must be validated. The scoring engine must base its score on a valid construct definition and handle unusual or bad-faith responses appropriately. Moreover, there must be a close match between the intended use of a system and key features of the scoring engine. At ETS, there are standard procedures and evaluation criteria for model building and validation: construct relevance and representation; association with human scores; association with other independent variables of interest; fairness of scores across subgroups; and impact and consequences of using automated scoring in operational settings. Because the specific features extracted by the e-rater engine are combined using a regression-based procedure that supports multiple scoring models, these models must also be validated.

Of particular interest in discussions of timed writing is the role of word count in AES systems. As Kobrin, Deng, and Shaw (2011) have noted, essay length has a significant, positive relationship to human-assigned essay scores. The association typically involves correlations above .60 but at or below .70. This relationship is not surprising given that words are needed to express thoughts and support persuasive essays. Length is not a meaningless indicator of writing ability; rather, it may be taken as an indicator of fluency. Shorter, lower-scoring responses often lack key features, such as development of supporting points, which contribute both to writing quality and to document length. Arguably, therefore, the association between document length and human scores reflects the ability of students to organize and regulate their writing processes efficiently. As long as an AES system measures features directly relevant to assessing writing quality, and does not rely on length as a proxy, this kind of association with length is both unavoidable and expected.

The design of Criterion, drawing upon the features built into the e-rater engine, is intended to help writers practice their writing, developing confidence and ultimately achieving

both fluency and effective voice, by providing real-time evaluation of their work in terms of grammar, usage and mechanics, features of style, and elements of essay structure. If we recognize that there are many paths to literacy, especially in digital environments (Black, 2009), then AES can and arguably should be viewed as but one tool to help students and their instructors along the way.

The Relationship of Automated Essay Scoring to Other Writing Assessment Systems

As both the educational measurement and writing assessment communities recognize, writing assessment should be validated in local settings. A general theoretical framework is suggested by Huot (1996) and Lynne (2004). Both argue that assessment must be site-based, locally-controlled, context-sensitive, rhetorically-based, assessable, meaningful, and ethical. Consistent with the criteria advanced in Cronbach (1975), local validation requires systematic inquiry focused on context-specific evaluation and improvement of instructional practices. The research reported here represents just such an evaluation and application.

Table 1 defines the six writing assessment systems at NJIT. Review of the systems, sponsors, purposes, content coverage, scoring, time allotted, and sequence suggest that the systems are themselves as complex as the construct they engage. Each has benefits but also drawbacks that must be taken into account to avoid claims that go beyond the uses for which each tool may validly be used. Given the tradeoffs, there may be much to gain by combining methods to take advantage of their different strengths. This approach allows one method to offset the disadvantages of another. The best ways to combine multiple assessment methods, however, is not clear in advance. For two years, we have been experimenting with each of these methods, focusing on determining what kind of information they provide, working to determine what uses they best support.

In the fall of 2010, the research team invited the entering first-year class at NJIT (N=1006) to participate in a rapid assessment so that students who were weak in the writing features covered by Criterion could be identified and writing program administrators could direct them to the university writing center for tutoring. Since the two submitted Criterion essays (n = 747) were timed at 45 minutes per persuasive prompt with an 800 word limit, we also asked students to submit, along with these two essays, samples that they had 48 hours to complete (n = 302), also written to college-level persuasive prompts. During that time, the students could draft and revise as they pleased and seek peer and instructor review. Seasoned faculty and instructional staff assigned essays scores on a 6-point Likert scale; resource constraints did not allow the 48 hour essays to be read twice. Table 2 presents the results of that study and the correlations between the three essays—two scored by Criterion, one scored by instructors using the Criterion rubric for a persuasive essay—and SAT Writing Section (SAT-W). While all correlations were statistically significant, the correlations are low-to-medium, with the first persuasive prompt correlating at somewhat higher levels with the SAT Writing and the 48 hour essay than the combined samples.

We also studied the relationship of Criterion and 48 hour essays to NJIT writing portfolios and to course grades at the end of the semester. Traditional, paper-based portfolios are designed to capture the student's best work in binders as cumulative demonstrations of their experiences with writing, reading, and critical analysis. At NJIT, writing portfolios are designed to yield information about program effectiveness (Middaugh, 2010) and are not intended to assess individual student performance. Portfolios are selected according to a sampling plan designed to yield a 95% confidence interval by using the smallest possible number of portfolios (Elliot, Briller, & Joshi, 2007). Following the writing, reading, and critical analysis experiences

outlined in the *Framework for Success in Postsecondary Writing* (CWPA, NCTE, WPA, 2011), the scoring rubric is designed to capture the variables of rhetorical knowledge, critical thinking, writing process, and knowledge of conventions. Portfolios are scored by two readers, with scores that differ by more than one point referred to a third reader. For fall 2010, a small sample of the traditional portfolios (n = 44) were read in order to infer reliability for the larger sample (n = 147). The scoring rubric asked raters to evaluate the portfolio on each variable with a 6-point Likert scale. Due to the complexity of the scoring task, the following weighted Kappa adjudicated ranges are lower than those found in timed essays: rhetorical knowledge ($\boldsymbol{K}$ = .63, p< 0.01); critical thinking ($\boldsymbol{K}$ = .47, p< 0.01); writing process ($\boldsymbol{K}$ = .69, p< 0.01); conventions ($\boldsymbol{K}$ = .62, p< 0.01); and holistic score ($\boldsymbol{K}$ = .62, p< 0.01). The relationship between the outcome variable (holistic score) and the predictor variables (rhetorical knowledge, critical thinking, writing process, and knowledge of conventions) is high: R = .87, R^2 = .76, F(4,142) = 110.16, p< 0.01. The holistic score is used to examine how well other measures agree with student performance as measured by the portfolio assessment.

While course grades are not often thought of as writing assessment systems, grades are nevertheless the most consequential and enduring assessment system used by schools. Willingham, Pollack, and Lewis (2002) have proposed a framework for understanding for possible sources of discrepancy in course-level grading, identifying such factors as content differences, specific skill assessment, components other than subject knowledge, individual differences, situational differences, and errors as sources of variance. Varying emphasis on any of these could result in differences between course grades and portfolios scores, especially at NJIT when portfolios are assessed independently (and often after) final grades are awarded.

With that background knowledge in mind, we turn to Table 3. The 48 hour essay, similar in construct domain to the two Criterion Essays, is moderately correlated with them. These correlations are higher than those obtained earlier in the semester, which is consistent with the fact that first year writing promotes a uniform approach towards the variables of writing ability covered by Criterion. The moderate relationships among those writing performances are also reflected in statistically significant, though lower, correlations with course grade. The low correlations between Criterion essay scores and the course grade can be explained both by the limitations of measurement imposed by taking only two writing samples and by the fact that the constructs directly measured by Criterion are a subset of the instructional goals of the course, designed to address the habits of mind of the *Framework for Success in Postsecondary Writing*. In fact, it is no surprise that the holistic portfolio score—an attempt to capture a much more robust view of writing performance—has no relationship with the Criterion essays, even though both have significant correlation with course grades. To gain insight into that lack of correlation, we must turn to the relationship between the course grade and the portfolio scores.

As Table 4 illustrates, there is a statistically significant moderate relationship between each of the traditional portfolio scores and the course grade; however, in the experimental sections of the course in which instructors used EPortfolios, there are no statistically significant relationships between the portfolio scores and the course grades. A review of Tables 1 – 4 reveals the importance of having multiple measures in writing assessment—as well as the importance of demonstrating wide construct coverage with those measures. Different writing assessment systems may tap different construct domains and only partially capture information about overall student performance. While examination of the relationship among systems is

necessary to reveal the best uses of each system, holding each system to the same claims of construct coverage is an error that may result in misuse that may, in turn, harm students.

Table 5 provides evidence of the usefulness of multiple systems of writing assessment. While the SAT-W predicted course grades with a statistically significant low correlation (R=.235), a model that combined the Criterion scores with the 48-hour essay predicted course grades at a much higher level (R=.475). For the case at hand, multiple samples of writing offer higher prediction rates than the combination of a multiple choice test and a single writing sample used in the SAT-W. Indeed, the inclusion of the SAT-W in the model increased the correlation only slightly (R=.476).

Further explication is in order. While the 48 hour essay yielded a statistically significant correlation (R=.375, p< 0.01) with the course grade that was higher than either the SAT-W or the combined Criterion scores, a regression model such as those shown in Table 5 could not be established because the 48-hour essay is a single variable. Nevertheless, it is important to note that the highest correlation with the course grade is produced from a sample that allowed students the most time to compose their submission; in fact, the correlation between the 48-hour essay and the final grade is higher than the .2 correlation reported by Peckham (2010) in his study of iMOAT, a system that allows extended time for essay submission. We cannot help but wonder, therefore, if the introduction of a second essay, written in a 48-hour time period, would yield a stronger model than those we have presently established. Yet, even if that model allowed comparable or higher prediction of the course grade, it would not be sustainable at NJIT in light of the inability to evaluate the existing 48-hour essay with more than one reader. For the specific institutional site at hand, an optimal model of prediction has thus been reached.

The Development of Localized Assessment Applications

It is important to have in place traditional measures that provide substantial construct coverage, such as portfolios; to encourage new forms of digital communication, it is equally important to experiment with innovative ways of capturing and assessing student performance. The availability of new tools creates new possibilities both for assessment and instruction, and it is advisable to consider how the tools can be put to use effectively rather than rejecting them because they do not fit existing practices or pedagogies. Whithaus (2006) provides a way forward by noting that data-driven investigations of how these systems are presently being used in postsecondary writing courses will be beneficial. Following this model, we turn now to the possibilities that arise from the combination of AES and digital literacy initiatives with existing writing assessment and instruction practices.

As Table 4 shows, while both the predictor and outcome variables are related in a statistically significant way to the course grade, an experiment with EPortfolios revealed both lower scores and lack of a statistically significant relationship. Examination of the low score on the variable asking readers to rate student performance on composing in electronic environments—the most elusive demand of the *Framework for Success in Postsecondary Writing*—revealed two types of EPortfolios in play. Some instructors used the EPortfolios as electronic filing cabinets and, as such, received the lower scores. Other instructors worked with their students to design web sites that required students to post documents, podcasts, and blogs to sections of Web sites they had designed to highlight their writing, reading, and critical thinking experiences, accompanied by the brief reflective statements advocated by White (2005). These EPortfolios (n = 17) received higher scores than the traditional portfolios in rhetorical knowledge (M = 9.1, SD = 2.3); in critical thinking (M = 8.47, SD = 2.57); in writing process (M = 8.47, SD

= 2.47); in knowledge of conventions (*M* = 8.76, *SD* = 2.25); in composing in multiple environments (*M* = 8.0, *SD* = 2.76); and in the holistic score (*M* = 9.06, *SD* = 2.49). There were, in fact, seven assessment systems under investigation, a reflection of the experimental exigency that arises in digital environments.

Because writing in digital environments provides occasions for experimentation and exploration, both Criterion and EPortfolios can be viewed, along with blogging and podcasting, as electronic tools. In fact, similar pedagogical aims in the development of these learning technologies are evident in an environment where students are encouraged to consider document design, information organization, and social networking as increasingly integral writing processes. Digital environments, it can be argued, present a much more complex framework for writing than print environments (Neal, 2011). Part of the change in intricacy derives from the technologies themselves. Electronic texts involve an ever expanding assortment of writing tools and programs, encapsulating nearly every stage writing, from concept generation, though data organization, to the design, presentation and even distribution of the final document. Such technological advances seem also to inform the progressive influence of communication theory on composition studies as the field begins to recognize elements of networking and multimedia support as fundamental to all digital modes of production (Rice & O'Gorman, 2008). Together, these two areas of development—digital communication technology and its theorization—are instrumental in transforming the study and practice of writing. In the future, as semantic technologies become more refined, allowing for the automation of additional components of the writing skills shown in Table 6, students will no doubt learn to reference an increasing number of tasks—improvement of sentence variety, for example—through software. Given these developments, it seems relatively easy to predict a deeper role for automated assessment

technologies in both instruction and assessment. Criterion exemplifies this trend, allowing students to practice and revise their writing before submitting it to examination by an instructor. The key issue in such practices is to determine how to use such tools to develop skills and facilitate success for writers attempting increasingly challenging writing tasks that might, without the digital technologies, have been too difficult.

Implications for Local Practice

For a specific institutional site, research in the use of AES yields benefits common to all educational research: an opportunity to identify methods leading to effective student learning. As part of New Jersey's science and technology university, all NJIT shareholders—alumni, administrators, instructors, students—embrace technology and are more than willing to entertain its applications. A field test of an AES on the Newark campus is therefore more likely than not to be welcomed, along with other digital applications such as an open source platform for course management or new blogging software. Each is distinct in its own way and of potential use for students. Indeed, mission fulfillment of NJIT—as judged by its regional accreditation agency, the Middle States Commission on Higher Education—relies on technological experimentation throughout the university, especially in student learning and its assessment. A framework of localism advocated by Cronbach (1975), suggests that, for institutions such as NJIT, research located at the intersection of technology and assessment of student learning is appropriate.

With the rise of digital writing frameworks, first-year writing in institutions such as NJIT find themselves in what Rice (2007) has called choral moments, pedagogical events that call into question many of the conventions surrounding print-based logic. As discontinuous as AES may appear in a print environment, it is strikingly continuous (and congruent) in the digital environment of NJIT (2010) in which the phrase "digital everyware" is part of a five-year

strategic plan intended to unify the university. For NJIT students, digital communication is part of professionalization and thus an important emphasis for the first year writing program. With the shift from print to digital environments, the digital medium, along with the tools and software needed to generate it, has become increasingly prominent. Transferred to digital media, the very concept of genre might be taught to students as both a form of response to exigence and as integral to design patterns that contribute to communication in complex contexts (Muller 2011). As we have seen in our study, digital technology's assessment capacity has broadened our understanding of writing experiences important to college and career success.

Is it a bridge too far to advance writing assessment by suggesting that it have a new relationship to digital pedagogy? Customary perspectives on writing and its evaluation have followed print-based conceptualizations of the rhetorical arts (Downs and Wardle 2007). Accordingly, assessment procedures attempt to control extraneous factors in the learning and evaluation context as strictly as possible, an effort that begins in most writing programs with an explicit call for evaluation standards and universal scoring tactics. Such endeavors to construct a stable scoring environment usually entail establishing well-defined, collectively accepted rubrics, as well as a shared understanding of different prose genres, number of assignments, and writing goals to be covered.

While AES technologies certainly do not eradicate the role of controlled context, they tend, by definition, to de-emphasize it. In these digital environments, students find themselves working with semantic technologies that incorporate assessment into the writing process itself, blurring the lines between formative and summative. The digital screen functions here less as a mode of authorial expression, as human reader scores on a rubric might; instead, students compose in an interactive medium in which an AES system such as Criterion becomes part of a

fluid environment where a machine score might be viewed as an invitation to revise instead of a judgment to be suffered. In a digital environment, terms such as rhetorical knowledge and writing assessment are re-imagined by students and instructors alike. As one first-year student recently noted in a course section emphasizing digital frameworks, audiences are static but networks are dynamic. The mental models underlying such a statement suggest that writing assessment, if it is to be truly responsive to new pedagogies, must itself be re-imagined.

Nevertheless, a critical stance to any such brave, new world includes concerns, and ours are similar to those reported by Perelman (2007) in his critique of the SAT-W. First, we wonder if our use of the 48 hour essay and Criterion will lead students to believe that knowledge of conventions is prerequisite to their experiments with print and digital exploration of rhetorical knowledge, critical thinking, experience with writing processes, and the ability to compose in multiple environments. In other words, we fear that we may have inadvertently achieved a 21st century surrogate of the error fixation that drove much of writing instruction in the early 20th century. Second, because the NJIT writing assessment system includes essays are machine scored, we fear that the machine will misjudge a writing feature and that students will be wrongly counseled. The suggested edits involved in word processing applications may take on the force of law in an assessment environment. Third, we fear that declining state budgets may result in an efficiency-minded administrator concluding that writing assessment can be undertaken by a machine. The next step, of course, might be to withdraw funding for first-year portfolio assessment, the system offering the most robust construct representation. Fourth, we fear that length of an essay, heft of a portfolio, or design of a web site may not, after all, be evidence of rhetorical power. Whether the regression analysis for AES depends too heavily on word count or the writing sample hyperlinked to a beautifully designed web portfolio lacks

critical thought is much the same. A system, no matter how technologically sophisticated or aesthetically designed, may remain just that and fail to justify anything beyond its own existence.

A Future for Automated Essay Scoring

What we have seen so far in this chapter is a detailed study of what automated essay scoring can accomplish now, based upon current technology and current assumptions about how it can be validated in local settings. It would, however, be a mistake to assume that technology will remain constant, or that future technologies will only measure the features captured in the present generation of AES systems. Current debates may be responding to a moment in time—in which the limited range of features shown in the shaded area of Table 6 have been incorporated into automated scoring technology—and in so doing, may risk forming too narrow a view of possibilities.

There is every reason to expect that future research will open up a wide range of features that provide much more direct information about many aspects of writing skill. For example, Deane and Quinlan (2010) briefly review new lines of research that are taking advantage of computer administration to collect keystroke logs, detailed records of student writing processes which may help us learn more about the significance of pauses first established by Perl (1979). Features derived from such logs may reflect different aspects of writing skill—such as longer pauses between sentences and paragraphs that may be likely to reflect planning and revision strategies—while a pattern in which writers display more and longer pauses for editing within words is likely to reflect general fluency of text production processes (Wengelin, 2007).

Consider, for example, some of the feature types for which automated measurement is currently available. Not all of these are currently used in AES systems like e-rater; but all are technically feasible: measurement of time on task; plagiarism detection; detection of off-topic

essays; detection of purely formulaic essay patterns such as the five paragraph essay; measurement of the complexity of essay organizational patterns; measurement of sentence variety; measurement of vocabulary sophistication; and detection of repetitive or stylistically awkward prose. Such features are indeed included in current automated scoring system either for construct relevance in predicting human scores or to detect cases to which the standard scoring models do not apply. But if we imagine an environment designed to encourage student writing, with automated feedback driven by an analysis of student responses, such features may have additional value as cues for feedback. Deployment of such features in an automated system holds the potential to allow additional construct representation of the writing skills shown in the non-shaded cells of Table 6, a representation that would yield more coverage of the habits of mind advocated in the *Framework for Success in Postsecondary Writing*.

The roles that writing assessment systems play depend on how they are integrated into the practices of teachers and students. If automated scoring is informed by enlightened classroom practice—and if automated features are integrated into effective practice in a thoughtful way—we will obtain new, digital forms of writing in which automated analysis encourages the instructional values favored by the writing community. Though AES is in a relatively early stage, fostering these values is the goal of the research we have reported.

References

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

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.

Baldwin, D. Fowles, M. & Livingston, S. (2005*). Guidelines for constructed response and other performance assessments*. Princeton, NJ: Educational Testing Service.

Black, R. W. (2009). Online fan fiction, global identifies, and imagination. *Research in the Teaching of English*, *43*, 397- 425.

Burstein, J. (2003). The e-rater® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-121). Hillsdale, NJ: Lawrence Erlbaum Associates.

Burstein, J., Chodorow, M., & Leacock C. (2004) Automated essay evaluation: The Criterion online writing service. *AI Magazine*, *25*, 27–36.

Council of Writing Program Administrators, National Council of Teachers of English, & National Writing Project. (2011). *Framework for success in postsecondary writing*. Retrieved from http://wpacouncil.org

Cronbach, L. J. (1975, February). Beyond the two disciplines of scientific psychology. *American Psychologist*, 116-127.

Deane, P.D., & Quinlan, T. (2010). What Automated Analyses of Corpora Can Tell Us About Writing Skills. *Journal of Writing Research*, *2*, 151-177.

Downs, D., & Wardle, E. (2007). Teaching about writing, righting misconceptions: (Re)envisioning "first-year composition" as "introduction to writing studies." *College Composition and Communica*tion, *58*, 552-584.

Elliot, N., Briller, V., & Joshi, K. (2007). Portfolio assessment: Quantification and community. *Journal of Writing Assessment, 3*, 5–30. Accessed http://www.journalofwritingassessment.org/

Huot, B. (1996). Towards a new theory of writing assessment. *College Composition and Communication*, *47*, 549-566.

Kobrin, J. L., Deng, H., & Shaw, E. J. (2011). The association between SAT prompt characteristics, response features, and essay scores. *Assessing Writing, 16*, 154-169.

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. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87-112). Hillsdale, NJ: Lawrence Erlbaum Associates.

Leacock, C., & Chodorow, M. (2003). C-rater: Scoring of short-answer questions. *Computers and the Humanities, 37*, 389–405.

Lynne, P. (2004). *Coming to terms: A theory of writing assessment*. Logan, UT: Utah State University Press.

Middaugh, M. F. (2010). *Planning and assessment in higher education: Demonstrating institutional effectiveness*. San Francisco, CA: Jossey-Bass.

Müller, K. (2011). Genre in the design space. *Computers and Composition*, *28*, 186-194.

Neal, M. R. (2011). *Writing assessment and the revolution in digital technologies*. New York, NY: Teachers College Press.

New Jersey Institute of Technology (2010). *Strategic plan, 2010 – 2015*. Accessed www. njit.edu

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48,* 238–243.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education 14*, 210–225.

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

Peckham, I. (2010). Online challenge vs. offline ACT. *College Composition and Communication*, *61*, 718-745.

Perl, S. (1979). The composing process of unskilled college writers, *Research in the Teaching of English*, *13*, 317-336.

Rice, J. (2007). *The rhetoric of cool: Composition studies and the new media*. Carbondale, IL: Southern Illinois University Press.

Rice, J., & O' Gorman (Eds.) (2008). *New media / new methods: The academic turn from literacy to electracy*. Anderson, SC: Parlor Press.

Rudner, L.M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning and Assessment, 4*(4). Available from http://www.jtla.org.

Sargeant, J., Wood, M. M., & Anderson, S. M. (2004). A human-computer collaborative approach to the marking of free text answers. *Proceedings of the 8th International CAA Conference* (pp. 361-370). Loughborough, UK: Loughborough University.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Singley, M. K., & Bennett, R. E. (1998). *Validation and extension of the mathematical expression response type: applications of schema theory to automatic scoring and item generation in mathematics* (GRE Board Professional Report No. 93-24P). Princeton, NJ: Educational Testing Service.

Wengelin, A. (2007). The word-level focus in text production by adults with reading and writing difficulties. In M. Torrance, L. van Waes & D. Galbraith (Eds.), *Writing and cognition: Research and applications* (pp. 67–82). Amsterdam, Netherlands: Elsevier.

White, E. M. (2005). The scoring of writing portfolios: Phase 2. *College Composition and Communication*, *56*, 581-600.

Whithaus, C. (2006). Always already: Automated essay scoring and grammar checkers in college writing courses. In P. E. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 166-176). Logan, UT: Utah State University Press.

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, *39*, 1-97.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0.* (ETS Research Report No. RR-08-62). Princeton, NJ: Educational Testing Service.

Table 1. Writing Assessment Systems in the ETS and NJIT Program of Research

System and Sponsor	Purpose	Content Coverage	Scoring and Time Allowed	Sequence
SAT Writing, College Board	Placement: Individual student placement into first-year writing	Essay: Develop and support point of view; follow conventions of standard written English Selected Response: improving sentences; identifying sentence errors; improving paragraphs	Essay: two human readers Selected Response: automated Time: 60 minutes	Before admission
48 Hour Essay, NJIT	Formative: Individual student rapid assessment; individual student end of semester assessment	Essay: topic development, insightfully, structure, sentence style, knowledge of conventions	One human reader, the student's instructor Time: 48 hours	Third through fifth week of semester; eleventh through fourteenth week of semester
Criterion, ETS	Formative: Individual student rapid assessment; individual student end of semester assessment	Essay: grammar, usage, mechanics, style, vocabulary, organization, and development	e-rater scoring engine Time: 45 minutes for each essay	Third through fifth week of semester; eleventh through fourteenth week of semester
Traditional Portfolios, NJIT	Summative: Writing program assessment	Portfolio: Rhetorical knowledge, critical thinking, writing process, conventions, holistic score	Two human readers, neither the student's instructor Time: 15 weeks	End of semester, after final grades posted
EPortfolios, NJIT	Summative: Writing program assessment	EPortfolio: Rhetorical knowledge, critical thinking, writing process, conventions, composing in electronic environments, holistic score	Two human readers, neither the student's instructor Time: 15 weeks	End of semester, after final grades posted
Course Grades, NJIT	Summative: Individual student end of semester assessment	Final Grade: Rhetorical knowledge, critical thinking, writing process, conventions, composing in electronic environments, information literacy	One human reader, the student's instructor Time: 15 weeks	End of semester

Table 2. Early Semester Fall 2010 Correlation of e-rater score for prompts A (n = 747), B (n = 745), and sum score for AB (n = 747), with SAT Writing Scores (n = 854) and the 48 Hour Essay (n = 302)

	A (*M* = 3.77, *SD* = 1.43)	*B* (*M* = 3.33, *SD* = 1.79)	*AB* (*M* = 7.09, *SD* = 2.93)	*M*	*SD*
SAT Writing	0.32	0.25	0.31	528	86.08
48 Hour Essay	0.26	0.17	0.24	3.82	1.09

Note: All correlations significant at the $p < 0.01$ level.

Table 3. Late Semester Fall 2010 Correlation of e-rater score for prompts C (n =328), D (n = 295), and sum score for CD (n = 290), with 48 Hour Essay (70), Holistic Portfolio Score (n = 147), and Course Grade (n = 824)

	C (*M* = 3.98, *SD* = 1.25)	*D* (*M* = 3.74, *SD* = 1.41)	*CD* (*M* = 7.69, *SD* = 2.4)	*M*	*SD*
48 Hour Essay	0.56**	0.43*	0.57**	3.64	1.8
Portfolio (Holistic Score)	0.16(ns)	0.18(ns)	0.2(ns)	7.98	2.01
Course Grade	0.29**	0.19**	0.28**	2.84	1.14

Note: * $p < 0.05$ level. ** $p < 0.01$ level

Table 4. Late Semester Fall 2010 Correlation of Course Grade (n =824, *M = 2.84, SD = 1.14)*) with Traditional (n = 147) and EPortfolio Scores (n = 48)

	Course Grade (*M* = 2.84, *SD* = 1.14) with Traditional Portfolio Scores	Course Grade (M = 2.84, *SD* =1.14) with EPortfolio Scores	*M*	*SD*
T. Rhetorical Knowledge	0.31	/	7.9	1.86
T. Critical Thinking	0.45	/	8.2	1.72
T. Writing Process	0.37	/	7.27	2.13
T. Conventions	0.43	/	7.99	1.9
T. Holistic Score	0.41	/	7.98	2.01
E. Rhetorical Knowledge	/	-0.06(*ns*)	6.58	3.17
E. Critical Thinking	/	0.07(*ns*)	6.21	3.11
E. Writing Process	/	0.05(*ns*)	6.33	3.05
E. Conventions	/	0.12(*ns*)	7.1	3.14
E. Electronic Environ.	/	-0.09(*ns*)	5.31	3.17
E. Holistic Score	/	-.000(*ns*)	6.44	3.36

Note: All correlations with Traditional Portfolio Scores are significant at the $p < 0.01$ level.

Table 5. Fall 2010 Prediction of Course Grade From Early Semester Measures

Models	Course Grade	Significance
1. SAT Writing	$R = .235$ $R^2 = .055$	$p < 0.01$
2. Criterion Essay Score 1 + Criterion Essay Score 2	$R = .321$ $R^2 = .103$	Essay 1 ($p < 0.01$) Essay 2 ($p = .902$ (*ns*))
3. Criterion Essay Score 1 + Criterion Essay Score 2 + 48 Hour Essay	$R = .475$ $R^2 = .225$	Essay 1 ($p < 0.01$) Essay 2 ($p < 0.01$) 48 Hour Essay ($p < 0.01$)
4. SAT Writing + Criterion Essay Score 1 + Criterion Essay Score 2 + 48 Hour Essay	$R = .476$ $R^2 = .226$	SAT Writing ($p = .202$(*ns*)) Essay 1($p < 0.5$) Essay 2 ($p < 0.01$) 48 Hour Essay ($p < 0.01$)

Note: All correlations with Traditional Portfolio Scores are significant at the $p < 0.01$ level.

Table 6: A Partial Analysis of Writing Skills. (Shaded cells represent skill types for which there are well-established methods of measurement using automated features)

	Expressive	**Interpretive**	**Deliberative**
	(Writing Quality)	(Ability to Evaluate Writing)	(Strategic control of the writing process)
Social Reasoning	Purpose, Voice, Tone	Sensitivity to Audience	Rhetorical strategies
Conceptual Reasoning	Evidence, Argumentation, Analysis	Critical stance toward content	Critical thinking strategies
Discourse Skills	Organization, Clarity, Relevance/Focus, Emphasis	Sensitivity to structural cues	Planning & revision strategies
Verbal Skills	Clarity, Precision of Wording, Sentence Variety, Style	Sensitivity to language	Strategies for word choice and editing
Print Skills		Sensitivity to print cues and conventions	Strategies for self-monitoring and copyediting