CHAPTER 6.

AUTOMATED ESSAY SCORING AND THE SEARCH FOR VALID WRITING ASSESSMENT

Andrew Klobucar, Paul Deane, Norbert Elliot, Chaitanya Ramineni, Perry Deess, and Alex Rudniy
New Jersey Institute of Technology and Educational Testing Service

In educational settings, assessment targets determine the need for local validation. Instructional improvement, for example, is validated by examining the relationship between curricular innovations and improvements in criterion measures such as course grades. In such cases, as both the educational measurement community (Cizek, 2008; Shepard, 2006) and the writing assessment community (Good, Osborne, and Birchfield, 2012; Huot, 1996; Lynne, 2004) recognize, assessments are most meaningful when they are site based, locally controlled, context sensitive, rhetorically informed, accountable, meaningful, and fair.

In the context of a first-year writing course, there are multiple reasons and occasions for measurement. Before a student enrolls, some information may be available and used for placement; but placement decisions are not perfect, and it is important to identify students who may require additional instructional support (Complete College America, 2012). At course completion, overall student performance must be assessed, both for the purposes of assigning course grades, and for purposes of program evaluation. Historically, New Jersey Institute of Technology (NJIT) has used measures such as the SAT Writing (SAT-W) for placement (Elliot, Deess, Rudniy, & Joshi, 2012). It has used human-scored writing samples allowing 48 hour completion to identify students for instructional support. Course grades are based upon teacher evaluation of student writing produced during the course. Student papers are also assembled into portfolios and human-scored on holistic and analytic rubrics for purposes of program evaluation. The availability of new technologies supports alternative approaches to scoring, such as Automated Essay Scoring (AES) systems, and alternative approaches to collecting samples of student work, such as the use of electronic portfolios (EPortfolios). Such innovations exemplify the 21st century emphasis on writing in digital environments.

DOI: https://doi.org/10.37514/PER-B.2012.0452.2.06

Because digital environments provide occasions for experimentation in teaching and assessing writing, both AES 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 information organization, document design, and social networking as increasingly integral to writing processes, products, and the audiences they serve. 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 of writing, from concept generation, through data organization, to the design, presentation and even distribution of the final document. Given these developments, it seems relatively easy to predict a deeper role for automated assessment technologies in both instruction and assessment. 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.

This chapter presents results from collaboration between NJIT and the Educational Testing Service (ETS). The focus of this collaboration is the Criterion® Online Writing Evaluation Service (Attali, 2004; Burstein, Chodorow, & Leacock, 2004), an integrated assessment and instructional system that collects writing samples and provides instant scores and annotated feedback focusing on grammar, usage and mechanics; style; and elements of essay structure.

Criterion exemplifies the trend toward writing in digital environments, and in particular, a movement toward making automated scoring and feedback available in such environments. Accordingly, systems have been developed for a variety of constructed-response tasks (Baldwin, Fowles & Livingston, 2005) including mathematical equations (Singley & Bennett, 1998), short written responses with well-defined correct answers (Leacock & Chodorow, 2003), and spoken responses (Xi, Higgins, Zechner, & Williamson, 2008). More than 12 different automated essay evaluation systems have been developed, including Project Essay Grade (Page, 1966, 1968, 2003), engine 5 (now available as Intelligent Essay Assessor from Pearson) from Knowledge Analysis Technologies™ (Landauer, Laham, & Foltz, 2003), Intelligent Essay Assessor (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 (Shermis & Burstein, 2003). Automated scoring can reproduce many of the advantages of multiple-choice scoring, including speed, consistency, transparent scoring logic, constant availability, and lower per-unit costs; because automated scoring is based on productive samples of student writing, it provides detailed performance-specific feedback (Shermis & Hammer, 2012).

The design of Criterion, drawing upon the features built into the e-rater engine, is intended to help writers achieve writing competency, develop confidence, and ultimately achieve fluency 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. It is, however, important to note that the value of automated methods to score writing is contested in many contexts. Concerns range from the signaling effect AES use sends about the general nature of composition studies to the specific impact of the technology on writing instruction and student learning (Bowen, Chingos, & McPherson, 2009). 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.

With the rise of digital writing frameworks, first-year writing programs 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. AES is strikingly continuous (and congruent) in the digital environment of NJIT 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 (Müller 2011).

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 contextual factors as strictly as possible, an effort that begins in most writing programs with an explicit call for evaluation standards and universal scoring tactics. Such efforts 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 do not eradicate the role of controlled context, they tend to de-emphasize it when integrated with other forms of digital communication. In digital environments, students find themselves working with technologies that incorporate assessment into the writing process itself. The digital screen functions here less as a mode of individual authorial expression, as human reader scores on a rubric might; instead, as subsequent research is demonstrating at NJIT, students compose in an interactive medium in which an AES system such as Criterion becomes part of a fluid environment where a machine score is 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 writing course emphasizing digital frameworks, audiences are static but networks are dynamic. The mental models underlying such a statement suggest that our concepts of writing must be reconsidered.

THE RELATIONSHIP OF AUTOMATED ESSAY SCORING TO OTHER WRITING ASSESSMENT SYSTEMS

However, we view such expansive possibilities, the immediate goal of assessment is to respond to existing needs and to improve current practices, often incrementally, and it is to such goals that we now turn. As we have already noted, several methods of writing assessment are at use at NJIT, including standardized tests, writing samples, course grades, and portfolio assessment of student work. These assessments differ in scope and applicability. Each has benefits but also drawbacks that must be considered to determine the uses for which each tool may validly be used. While, for instance, portfolios address the fullest possible range of the target domain of writing that can be demonstrated in a first-year course, other assessments such as the SAT-W, the 48 hour essay, and Criterion address a subset of that target domain. While timed writing is not part of the portfolios, the command of construct coverage associated with the brief essay, especially knowledge of conventions, is significant in establishing course grades. 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. Since 2009, 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 case of existing measures, a great deal already is known. SAT-W as a measure of writing skill has been discussed extensively elsewhere (Bowen, Chingos, & McPherson, 2009), and need not be discussed in detail here. It is a useful, though partial, indicator of writing competency for purposes of admission or placement. The 48 hour human-scored writing samples are typical instances of the use of direct writing assessment in writing program assessment (Adler-Kassner & O'Neill, 2010; Huot, 2002). More attention should be focused on the two end-of-course measures: traditional, paper-based portfolios and course grades.

Traditional, paper-based portfolios are designed to provide cumulative demonstrations of student 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.

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 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.

There are two new measures we are currently exploring: use of EPortfolios and AES. In the study reported in this chapter, implementation of EPortfolios was in its first year, and too few electronic portfolios were available to support a meaningful comparison with existing measures or with AES. We therefore focused on AES, and in particular, on the use of Criterion to provide embedded assessment within the writing course.

In the case of AES, the usefulness of the assessment is judged by its ability to reliably assess student writing according to a defined construct model of writing (Shermis & Hamner, 2012). 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, these models must also be validated. These kinds of validations have been done on general populations as part of the development of Criterion (Attali, 2004; Burstein, Chodorow & Leacock, 2004). However, the place of the construct that Criterion measures within a curriculum, in tandem with the role it plays within a local setting, requires validation within an institution. We are actively engaged in research to train and validate e-rater models specifically for the NJIT population, but in the study reported here, we use off-the-shelf Criterion prompts and e-rater scoring models. The results we report should therefore be viewed as establishing a baseline of Criterion performance in the context and use described, and not as establishing a ceiling.

DESIGN OF THE 2010 STUDY

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 = 603) 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 = 300), 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 precluded having the 48 hour essays read twice.

In addition to the writing samples, course grades were collected for all students, and a random sample of traditional paper portfolios was scored (N=135). A subset of these portfolios (n = 44) were read twice in order to infer reliability for the larger sample. Both trait scores and a holistic score were collected. The holistic score was selected as the most directly parallel for purposes of comparing the paper portfolios with other measures. As a follow-up measure, a second round of e-rater scores, was collected at the end of the semester, but the total number of students participating (N = 249) was relatively low, and the intersection between this group and the group of students for whom traditional portfolios were collected was even smaller (N = 57). We therefore excluded the December Criterion administration from the analysis presented below.

RESULTS AND DISCUSSION

The dataset we analyze thus contains SAT-W scores, scores on the two automatically-scored essays in Criterion, which we considered both separately and summed, scores on the 48 hour human-scored essays, course grades, and holistic traditional portfolio scores. Descriptive statistics for these measures can be found in Table 1.

Table 1. Descriptive statistics for all writing performance measures and end-of-course grades

Measure	**N**	**M (SD)**	**(Min, Max)**
Prior to the semester			
SAT Writing	735	526 (82)	300, 800
At the beginning of the semester			
Criterion essay 1	603	4.17 (0.85)	1,6
Criterion essay 2	603	4.08 (0.94)	1,6
Combined Criterion score	603	8.25 (1.64)	2,12
The 48 hour essay	300	3.85 (1.06)	1,6
At the end of the semester			
Combined Criterion score	273	8.03 (1.97)	2,12
Traditional Portfolio	135	8.13 (1.90)	2,12
EPortfolio	44	7.02 (2.86)	2,12
Grades	736	2.95 (1.04)	0,4

Traditional portfolio scoring was performed using standard NJIT methodology and rubrics. Due to the complexity of the task, the following weighted Kappa adjudicated ranges are lower than those found in timed essays: rhetorical knowledge ($K = .63$, $p < 0.01$); critical thinking ($K = .47$, $p < 0.01$); writing process ($K = .7$, $p < 0.01$); conventions (K = .63, $p < 0.01$); and holistic score ($K = .62$, $p < 0.01$). However, the relationship between the outcome variable (holistic portfolio 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. We therefore were confident in using the holistic portfolio scores as a criterion measure.

Correlations between portfolio trait scores and course grade were in the moderate range (.35-.5). The correlation between the holistic portfolio score and course grade was at the high end of that range (.43). Grades are subject to many additional influences above and beyond writing competency (Willingham, Pollack & Lewis, 2002), and so the size of these correlations is in the expected range, comparable to those observed in earlier years of portfolio assessment with NJIT students (Elliot, Briller, & Johsi, 2007; Elliot, Deess, Rudniy & Johsi, 2012).

Correlations between SAT Writing scores, Criterion essay scores, traditional portfolio scores, and course grades are shown in Table 2. Correlations between the timed writing prompts fall in the moderate range (.29-.41). Correlations between these measures and the end-of course measures fell in a similar range (.24-.43 for grades, .32-43 for traditional portfolios.) The e-rater correlations are slightly lower than the correlations for the 48 hour essay, but equal to or higher than correlations for SAT Writing.

As an embedded assessment, Criterion can be use as an early warning system for instructors and their students. While 10 to 15 percent of admitted students are traditionally placed in basic writing at NJIT, a combined criterion score of 6 (15.6 cumulative percent of score frequency) was used as an early warning score so instructors could identify potentially at-risk students for writing center and tutoring help. Of the 93 students earning scores of 6 or below early in the semester, only 12 students (13 percent) received a grade of D or F; that is, 16 percent received a grade of C, 17 percent received a grade of C+, 30 percent received a grade of B, 10 percent received a grade of B+, and 14 percent received a grade of A. Such student success suggests the value of Criterion for embedded assessment and early warning. Because Criterion was primarily at the beginning of the semester in the fall of 2010, decline in student use is clear as the number of submissions declined from 603 combined scores to 273 combined scores at the end of the semester. Emphasis on using Criterion throughout the semester remains a challenge.

Table 2 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. The moderate, statistically significant relationship of the target domain of Criterion and that of the 48 hour essay provide convergent validity evidence that the two assessments—similar to the SAT-W—are different mea-

Table 2. Correlations between writing performance measures from prior to (or beginning-of) semester and end of semester portfolio measures and course grades, with number of student submissions

	SAT Writing	Criterion Essay 1	Criterion Essay 2	Combined Criterion score	The 48 hour Essay	Traditional Portfolio
SAT writing	1					
Criterion Essay 1	0.42 (591)	1				
Criterion Essay 2	0.34 (591)	0.68 (603)	1			
Combined Criterion Score	0.41 (591)	0.91 (603)	0.93 (603)	1		
The 48 hour Essay	0.41 (296)	0.31 (274)	0.23 (274)	0.29 (274)	1	
Traditional Portfolio	0.40 (135)	0.42 (116)	0.32 (116)	0.39 (116)	0.43 (56)	1
Grades	0.25 (720)	0.29 (595)	0.24 (595)	0.29 (595)	0.35 (296)	0.43 (135)

Note. All correlations significant at the p < 0.01 level. EPortfolio not included because of the small N. EPortfolio correlations with SAT-W and E-rater scores are > .25, but not significant since for N=45, only correlations > .288 will be significant at the .05 level.

sures of a related trait (Brennan, 2006). Indeed, the relatively slightly lower correlations between Criterion essay scores and the end-of-course measures may be related to the fact that the constructs directly measured by Criterion are a subset of the instructional goals of the course, designed to address the writing and critical analysis experiences of the *Framework for Success in Postsecondary Writing*, and so may be necessary, but not sufficient, to achieve success in the course.

Regression analyses shown in Table 3 provide further evidence of the relation among the timed essays and their ability to predict end-of-course scores. Since the intended use of e-rater scores was to substitute for the 48 hour essay in identification of students who might be in need of instructional support, we examine the effects of using the e-rater scores and the 48 hour essay scores both alone and in combination with SAT Writing scores. Corresponding to the moderate correlations observed in Table 2, we observe low R^2 values, but relatively small differences between the three predictors. The 48 hour essay performed better than the combined e-rater scores, which performed better than the SAT Writing prompt. However, the differences were relatively small. If we combine the SAT Writing score with the Criterion essay scores, the resulting model exceeds the performance of the 48 hour essay (R^2 = .20 vs. .17) in predicting traditional portfolio scores, and is only slightly less effective at predicting course grades (R^2=.10 vs. .12). Combining the 48 hour essay score with SAT Writing improves prediction of grades slightly (R^2 = .14 instead of .12), but when applied to traditional portfolio scores, fails to yield a model in which the 48 hour essay is a significant predictor.

Table 3. Prediction of end of semester portfolio scores and course grades using prior to (and/or beginning of) semester writing performance measures

Model	RSquare for the outcome	
	Traditional Portfolio	**Grades**
SAT Writing	0.15	0.06
Combined Criterion score	0.14	0.08
The 48 hour essay	0.17	0.12
SAT Writing + Combined Criterion score	0.20	0.10
SAT Writing + The 48 hour essay	- *	0.14

**RSquare = 0.31, but model rejected since regression coefficient for the 48 hour essay was not significant. The N=56 for this model is very small. All other predictors significant p < 0.01 level.*

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. These results suggest that although the 48 hour essay scores are a better predictor of end-of-course performance than 2 45-minute essay scores, they are only marginally better—and have the disadvantage of requiring human scoring of more than a thousand essays within a very short timeframe. Since the purpose of assessment is to identify students in need of instructional support, a purely formative use, the case for using e-rater scores instead of 48 hour essays is relatively strong based on grounds of practicality, subject to further validation and evaluation.

IMPLICATIONS FOR LOCAL PRACTICE

While it is important to have in place traditional measures that provide substantial construct coverage, such as portfolios, it is equally important to experiment with innovative ways of capturing and assessing student performance in order to encourage new forms of digital communication. For institutions such as NJIT, research located at the intersection of technology and assessment of student learning is appropriate. Indeed, mission fulfillment for 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. 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. It is in this spirit that we have undertaken the work reported in this study. However, it would be a mistake to focus solely on the results of a single study, or even on the possibilities for using a particular AES tool such as Criterion, or to imagine that innovations will long be restricted in their scope. The possibilities for new forms of local practice are inherent in the spread of digital communications technology, and the most important role that local writing communities can play in this process is to help to shape it.

The availability of new tools such as Criterion creates new possibilities both for assessment and instruction, and it is advisable to consider how these tools can be put to effective use. 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. In a similar fashion, Neal (2011) has provided a direction for experimentation with digital frameworks for writing instruction and assessment by focusing on hypertext (connections in EPortfolios), hypermedia (multimodal composition), and hyperattention (information processing). Together, these two areas of development—digital communication technology and its theorization—are instrumental in transforming the study and practice of writing.

Nevertheless, a critical stance to any such brave, new world includes concerns, and ours are similar to those reported by Perelman (2005) in his critique of the SAT-W. First, at NJIT 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 must be on guard against 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 that are machine scored, we guard against the possibility that the machine will misjudge a writing feature and that students will be wrongly counseled. As ever, machines make good tools, but terrible masters. Third, we are alert to the possibility that declining state budgets may result in an efficiency-minded administrator concluding that the whole of writing assessment can be accomplished through machine scoring. 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 must never forget that surface features such as the length of an essay, heft of a portfolio, or design of a web site are not proof of rhetorical power. There is very little difference between an AES system that relies too heavily on word count and the instructor who gives high scores to a beautifully designed web portfolio that lacks critical thought in the documents uploaded to it. A system, no matter how technologically sophisticated or visually well-designed, may fail to justify anything beyond its own existence.

What we have seen thus far is a baseline study of the role that AES can play at a specific institutional site, based upon current technology and current assumptions about how it can be validated in local settings. It would be a mistake to assume that technology will remain constant, or that future technologies will only measure features captured in the present generation of AES systems. 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.

Consider some of the features for which automated measurement is currently available, such as plagiarism detection; detection of off-topic essays; detection of purely formulaic essay patterns; measurement of organizational complexity; measurement of sentence variety; measurement of vocabulary sophistication; and detection of repetitive or stylistically awkward prose. Such features may be useful for scoring. 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 that is fully integrated with the writing process. As technology advances, it may be possible to deploy features that that support effective writing shown in the non-shaded cells of Table 4, a representation that would yield more coverage of the writing and critical analysis experiences advocated in the *Framework for Success in Postsecondary Writing*. (See Deane, 2011, for a more detailed outline of these ideas.) In the future, as linguistic technologies become more refined, students

Table 4. A partial analysis of writing skills

	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

Note. Shaded cells represent skill types for which there are well-established methods of measurement using automated features.

will no doubt learn to reference an increasing number of tasks—improvement of sentence variety, for example—through software (Deane, Quinlan, & Kostin, 2011).

More generally, we would argue, it is very likely that current debates are responding to a moment in time—in which the limited range of features shown in the shaded area of Table 4 have been incorporated into automated scoring technology—and in so doing, may risk forming too narrow a view of possibilities. 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.

NOTES

[1] 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. 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 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, an association with length is both unavoidable and expected.

2. While the work is in a fairly early stage, differences in instructor practice are already revealing, and underscore the importance of (re)centering rhetorical frameworks in digital environments (Neal, 2011). Analysis of the contents of the portfolios revealed that some instructors used the EPortfolios as electronic filing cabinets. 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 average scores than traditional portfolios when scored to the same rubric, though the number of cases is too small to draw any firm conclusions at the present time.

REFERENCES

Adler-Kassner, L., & O'Neill, P. (2010). *Reframing writing assessment to improve teaching and learning.* Logan, Utah: Utah State University Press.

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). Retrieved 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.

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (2009). *Crossing the finish line: Completing college at America's public universities. P*rinceton, NJ: Princeton University Press.

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

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.

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

Cizek, G. J. (2008). Assessing Educational measurement: Ovations, omissions, opportunities. [Review of *Educational measurement, 4*th ed., by R. L. Brennan (Ed.).] *Educational Researcher, 37,* 96-100.

Complete College America (2012). R*emediation: Higher education's bridge to nowhere.* Washington, DC: Complete College America. Retrieved from http://www.completecollege.org/docs/CCA-Remediation-final.pdf

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

Deane, P. (2011). *Writing assessment and cognition.* (ETS Research Report RR-11-14). Princeton, NJ: Educational Testing Service.

Deane, P., Quinlan, T., & Kostin, I. (2011). *Automated scoring within a developmental, cognitive model of writing proficiency* (No. RR-11-16). Princeton, NJ: Educational Testing Service.

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

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

Elliot, N., Deess, P., Rudniy, A., & Johsi, K. (2012). Placement of students into first-year writing courses. *Research in the Teaching of English, 46, 2*85-313.

Good, J. M., Osborne, K., & Birchfield, K. (2012). Placing data in the hands of discipline-specific decision-makers: Campus-wide writing program assessment. *Assessing Writing, 17, 140-149.*

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

Huot, B (2002). *(Re)Articulating writing assessment for teaching and learning.* Logan, Utah: Utah State University Press.

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.

Leacock, C., & Chodorow, M. (2003). C-rater: Scoring of short-answer questions. *Computers and the Humanities, 37, 3*89–405.

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.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48, 2*38–243.

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.

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

Perelman, L. (2005, May 29). New SAT: Write long, badly and prosper. *Los Angeles Times.* Retrieved from http://articles.latimes.com/2005/may/29/opinion/oe-perelman29

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

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

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623-646). Westport, CT: American Council on Education and Praeger.

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

Shermis, M. D., & Hammer, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. Paper presented at the annual meeting of the National Council of Measurement in Education, Vancouver, BC, Canada.

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.

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 RR-08-62). Princeton, NJ: Educational Testing Service.