13

AUTOMATED ESSAY GRADING IN THE SOCIOLOGY CLASSROOM

Finding Common Ground

Edward Brent and Martha Townsend

OVERVIEW

This chapter describes an effort by one author, a sociologist, to introduce automated essay grading in the classroom, and the concerns raised by the other author, the director of a campuswide writing program, in evaluating the grading scheme for fulfillment of a writing-intensive (WI) requirement. Brent provides an overview of existing automated essay-grading programs, pointing out the ways these programs do not meet his needs for evaluating students' understanding of sociology content. Then he describes the program he developed to meet those needs along with an assessment of the program's use with six hundred students over three semesters. Townsend provides a brief overview of the campus's twenty-year-old WI graduation requirement and illustrates concerns raised by the Campus Writing Board when Brent's course, employing the machine-graded system, was proposed for designation as WI. In this point-counterpoint chapter, the coauthors highlight areas of concordance and disagreement between the individual professor's use of machine-graded writing and the established writing program's expectations.

INTEGRATING AUTOMATED ESSAY GRADING INTO AN INTRODUCTORY SOCIOLOGY COURSE: BRENT'S PERSPECTIVE

I have taught introductory sociology for many years to classes of 150 to 250 students each semester. By necessity, my course has relied almost exclusively on in-class multiple-choice tests for evaluation. Students often express frustration at taking such tests, and I find it very hard to measure higher-level reasoning on these tests. My objective is to incorporate more writing into this large-enrollment course despite limited TA resources.

Using writing for learning and assessment offers a number of advantages over multiple-choice tests (Bennett and Ward 1993). Essays are

more "authentic" than multiple-choice tests because they "present test-takers with tasks more similar to those in the actual educational or job settings" (Yang, Buckendahl, and Juszkiewicz 2001). Essays permit students to demonstrate higher-order thinking skills such as analysis and synthesis (Rudner and Gagne 2001), requiring students to construct arguments, recall information, make connections, and support their positions (Shermis and Burstein 2003).

However, grading essays is expensive and time-consuming (Rudner and Gagne 2001). Feedback is often delayed, limited in scope, and of poor quality (Yang, Buckendahl, and Juszkiewicz 2001). Adding significant writing assignments to this large-enrollment introductory course required a new, more cost-effective strategy, so investigating automated essay grading programs seemed worthwhile.

Automated Essay-Grading Programs

A number of commercially available essay-grading programs are used in some very high-profile applications. Several large-scale assessment programs now include one or more measures based on writing, including "the Graduate Management Admissions Test (GMAT), the Test of English as a Foreign Language (TOEFL), the Graduate Record Examination (GRE), Professional Assessments for Beginning Teachers (Praxis), the College Board's Scholastic Assessment Test II Writing Test and Advanced Placement (AP) exam, and the College-Level Examination Program (CLEP) English and writing tests" (Burstein 2003, 113). Many of these tests also have students submit essays by computer, including the GMAT, TOEFL, GRE, and Praxis, making the use of automatic-scoring programs feasible for those tests. Commercially available essay-grading programs used in these tests include the Intelligent Essay Assessor, the erater, developed by Burstein and her colleagues at the Educational Testing Service, and the IntelliMetric program .

Some of these programs employ a statistical approach for developing and assessing the automated-grading model. In each case human graders must first grade many (usually several hundred) essays. Those overall grades are then used as the "gold standard" to fit or "train" statistical models predicting scores assigned by human graders from features of essays measured by the programs (Yang, Buckendahl, and Juszkiewicz 2001). Once trained, the resulting model can then be used to assign grades to papers in the test set without using human graders.

Other programs for automated essay grading take a rule-based or knowledge-based approach; in these, expert knowledge provides the

standard for assessing student performance. One or more experts creates a knowledge base for the content area along with a grading rubric indicating the kinds of knowledge and reasoning students should display. Student essays are examined for evidence of such knowledge, with better scores being given to students whose writing most closely expresses the expert knowledge. A rule-based or knowledge-based program can be tested on a much smaller number of cases, thereby reducing development costs. This approach obviously requires an expert to explicitly determine the knowledge-based criteria

Concerns and Standards for Essay-Grading Programs

Automated essay-grading programs appear to offer a number of advantages over manual grading of essays. They are much faster than human readers, often being able to score essays in only a second or two. Hundreds or even thousands of essays can be graded very quickly and efficiently, with less cost than manual grading, and the scores are immediately provided to students (Rudner and Gagne 2001; Yang, Buckendahl, and Juszkiewicz 2001). However, a number of criteria must be considered in deciding whether and which automated-grading program to use, including the nature of the writing task, cost-effectiveness, an appropriate standard for assessing writing, and the quality of feedback provided to students.

The Writing Task

Statistical programs work for standardized writing assessment, in which the "mechanics" of writing—spelling, punctuation, subject-verb agreement, noun-pronoun agreement, and the like—are being scored, as opposed to substantive, discipline-based knowledge. Essays with very general topics often have few or no "content" constraints, in order to permit students from a wide range of backgrounds to answer the given prompt. They typically address broad questions having no right or wrong answer while giving writers an opportunity to construct an argument, organize their thoughts, and show that they can reason about the problem. In this kind of assessment, mechanics along with some organizational and reasoning abilities are more important than discipline-based content. For such tasks, statistical programs that assign grades based on the grades assigned to similar papers by human graders may be appropriate.

In contrast, in most writing tasks for discipline-based courses dealing with substantive knowledge in the field—whether they be term papers,

shorter formal or informal assignments, or answers to tests—there is greater emphasis on content. Mechanical skills such as spelling and punctuation are secondary to being able to construct an argument, reason in accepted ways, and understand specific content. Writing tasks for discipline-based courses are usually designed to assess students' understanding and knowledge of the substantive domain of the course, along with their ability to perform the kinds of higher-order reasoning that are important for that discipline. For example, in sociology we want students to be able to develop and understand a causal argument, to recognize specific theories and the concepts and proponents associated with them, to identify examples of a concept, to interpret specific events from different theoretical perspectives, and to understand and critique the methods used in studies. The ability of students to construct arguments using these forms of reasoning and specific substantive knowledge is best measured with rule-based programs.

Cost-Effectiveness

We would expect automated essay-grading programs that can grade literally hundreds or thousands of papers an hour without human intervention to cost much less than grading those same essays with human graders. However, the cost and time required to develop machine-scoring systems can be prohibitive (Yang, Buckendahl, and Juszkiewicz 2001). In an actual trial of machine grading, Palmer, Williams, and Dreher (2002) found the cost of machine grading to greatly exceed the cost of grading by human graders for a few hundred essays due to high up-front development costs. Automated grading is most cost-effective for large numbers of essays where minimal costs are required for training the program and the users pay a one-time fee for use of the program. Commercially available automated-grading programs are usually not cost-effective for small classes and nonstandardized teaching and assessment.

The economics of statistical approaches and rule-based approaches are somewhat different. The statistically based programs generally require that a few hundred student essays be graded by competent human graders, then those data are used to estimate parameters of the regression equations for the model. In contrast, knowledge-based approaches require that an expert in the discipline specify the correct knowledge. In this case only a few essays need to be graded to test the program's ability to detect information in student essays correctly. Hence, rule-based programs are more likely to be cost-effective even for moderately large classes.

An Appropriate Standard for Assessing Performance

Statistical programs for automated essay grading use the grades assigned to similar papers by human graders as their standard for judging essays. However, "the correlation of human ratings on (essays) is typically only .70 .75" (Rudner and Gagné 2001), and exact agreement among human judges is often in the 50 percent to 60 percent range. "Thus, correlating with human raters as well as human raters correlate with each other is not a very high, nor very meaningful, standard" (2). A more appropriate standard for judging writing is whether it displays important features we expect in good writing rather than whether it displays indirect measures that correlate with human readers' scores (Page 1966; Page and Petersen 1995) or whether it matches documents having similar scores (Landauer et al.1997). The important issue is not consistency with human graders but the validity of the scores. Hence, rule-based essay-grading programs provide a better standard for judging student work (Klein et al. 2001).

Quality of Feedback

Essay-grading programs based on statistical modeling have often been criticized for being unable to provide good feedback (Kukich 2000), giving students little or no advice on how to improve their scores. Those programs sometimes produce only a single summary grade, or at most only a few summary measures. Poor feedback may be a fundamental weakness of statistical approaches because they are based on complex patterns of statistical relationships that may be hard to interpret and indeed may have little meaning to either readers or writers. Also, most currently available programs for automated essay scoring are proprietary commercial systems, and their algorithms are treated as trade secrets, described only in generalities. We do not know, for example, the specific variables used in any model nor their weights in predicting the overall score (Rudner and Gagné 2001; Shermis and Burstein 2003).

In contrast, in rule-based programs the criteria are determined based on experts' knowledge; criteria are chosen because they reflect meaningful knowledge the writer should be able to display. In many cases, rule-based programs have an explicit rubric indicating what features should be present and how many points are assigned for each. For this reason, rule-based programs like the Qualrus-based SAGrader program are able to provide very explicit and detailed feedback to students and instructors that clearly states what they did right (or wrong) and how students can improve their grades.

For all these reasons, I chose to use a rule-based program for automated essay grading rather than a statistical one. For this purpose, my colleagues and I developed our own essay-grading program, SAGrader. This program builds upon a general-purpose qualitative analysis program, Qualrus, which we also developed and which is widely used in both industry and academia for analyzing unstructured data. Both of these programs are available commercially from Idea Works, Inc.

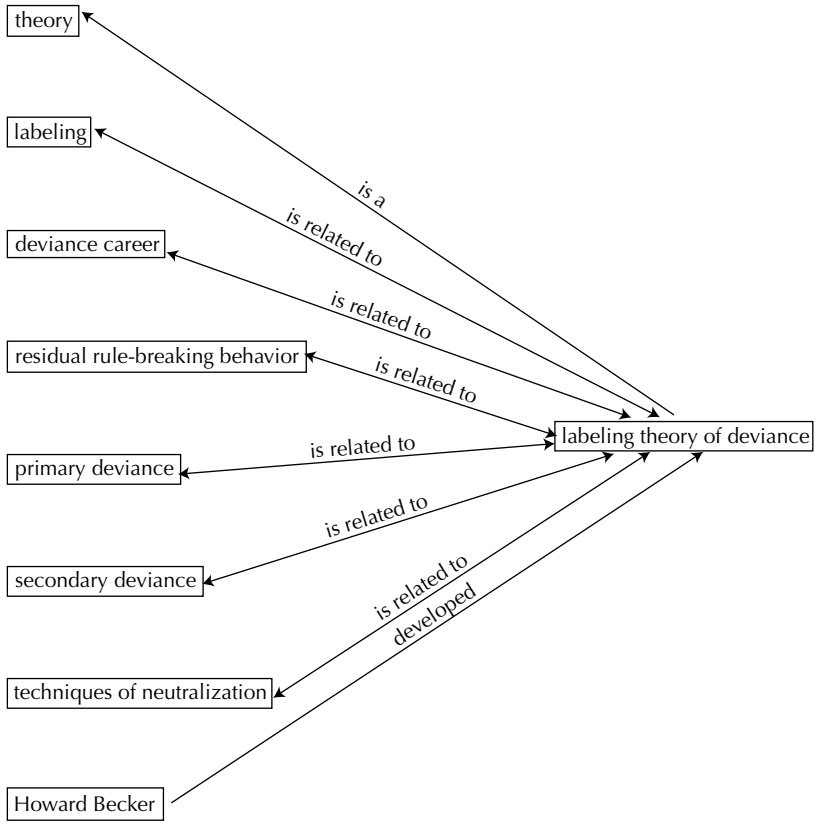
A Substantive-Based Approach to Automated Essay Grading

My objective is to assess students' discipline-based substantive knowledge and reasoning by having them write several brief focused papers addressing specific substantive objectives. This approach emphasizes substantive content over writing skills and focuses on students' knowledge of sociological concepts, theories, and methods; the approach also emphasizes students' ability to use this knowledge to reason sociologically about the world around them. The Qualrus-based SAGrader program developed for this course expresses substantive knowledge as a semantic network linking key concepts, theories, authors, studies, and findings from sociology. The structure of that knowledge base gives the program relevant information that can be used to help identify student misunderstandings and generate individualized student feedback. It then uses rudimentary natural-language processing to recognize key terms and phrases in text that reflect relevant elements such as concepts or theories based on those available in the semantic network for each chapter. It uses the grading rubric (and the relatively structured assignment) to create a template describing the rhetorical objectives for the writing assignment. By comparing each student's written input with the set of requirements for the assignment, the program is able to grade essays. For example, in the program's substantive knowledge base, the labeling theory of deviance has concepts and theorists related to it as shown in figure 1 (next page).

One writing assignment asks students to briefly describe one theory of deviance. If they chose this theory, the program would look for these concepts and theorists, giving students points when they correctly identify concepts and theorists associated with this theory, subtracting points when they leave out important items or include incorrect concepts or theorists from other theories. The grading rubric specifies how many points are assigned for each element.

For the writing-intensive course (WI) there is one written paper (fifteen pages long) plus four two-page writing exercises requiring students

FIGURE 1



to address specific issues. The two-page exercises employ common forms of sociological reasoning to address a range of specific substantive concepts and perspectives. For example, the second assignment asks students to write about crime and theories of deviance (see figure 2).

The four writing exercises and term paper demand increasingly higher levels of reasoning as the student progresses, with later exercises requiring students to interpret a description of a community sociologically and to use sociological concepts and perspectives to describe and understand their own families. In the future I hope to provide additional writing exercises and permit students to select which ones they will do.

To submit their work, students enter WebCT and use a hyperlink on the syllabus to open the submission form in their Internet browser. There they enter their names and student numbers, then paste their

FIGURE 2

Writing Exercise 2: Crime and Theories of Deviance

20 points, 2 drafts, 2 pages each time, first draft reviewed by computer (optionally by TAs as well), second draft reviewed by TAs. The final score is the weighted average of first (1/3) and second drafts (2/3).

Assignment: Select a type of crime discussed in the chapter on deviance and social control. Briefly describe this type of crime, give examples of it, and indicate any other types of crime it might be closely related to. Then take one of the theories of deviance discussed in this same chapter, briefly summarize the theory, and discuss how well that theory can account for the type of crime you have chosen. Your answer should identify the theory, one or more proponents of the theory, and four or more key concepts from that theory.

Learning Objectives: To become familiar with the components of theories, concepts, proponents, and how those theories can be used to explain specific phenomena.

Substantive Topics: Theories of deviance, types of crime.

papers into a text box. Once this is done, they press a "submit" button at the bottom of the page and the paper is sent to the Qualrus server, where it is graded and detailed feedback is displayed on a second Web page. The time between submitting the paper and receiving detailed feedback is usually two seconds or less, depending on network transfer speeds. A sample of student writing and the feedback the program provides on the deviance assignment is shown in figure 3 (next page).

The program feedback is detailed and specific, showing students not only their scores for specific components but also specific ways they can improve their score.

Strengths and Weaknesses of This Approach

This essay-grading program offers the promise of speeding up the grading process, reducing costs, and giving students the opportunity to write in a large lecture class. However, the program has limits and, to be successful, must be carefully integrated into the course to overcome those limits.

The approach has the advantage of providing very explicit criteria for judging the essay that are consistent with the learning objectives of the course. Consequently, the program can provide students with a detailed breakdown of credit received and missed based on their answers along with detailed feedback of how their answers do and do not meet the assignment objectives. Since the feedback is nearly instantaneous, it can provide a much better learning environment for students than they experience when human graders take days or weeks to grade their papers.

Each assignment combines specific substantive content represented in a semantic network (such as descriptions of a theory or a study)

FIGURE 3

SAMPLE ANSWER (Abbreviated)
Qualrus Grading Server
The Nature of Crime
by: Samuel Adams

Crime affects everyone in the United States. There are many types of crime. I will focus on white-collar crimes such as crimes committed by someone who knowingly pollutes a stream from runoff from their hog farm. White collar crime can be understand in terms of labeling theory. Labeling Theory labels deviants. The labeling has two different types. One is primary deviance and the second is secondary deviance. Primary deviance is goes undetected by authorities and secondary deviance is known by the authorities and they accept that title. Once people get labeled as a deviance they have a hard time getting rid of the title.

SAMPLE FEEDBACK (Abbreviated)
Qualrus Grading Server
Essay Grader
Copyright © 2002, Idea Works, Inc.
Student Feedback for Samuel Adams

Number of Concepts from the Chapter: +2 out of 7 points
You appear to have included only 5 concepts from this chapter, instead of the 10 required. The concepts you included that are from this chapter are crime, labeling, primary deviance, secondary deviance, and time . . .

Overall Treatment of Theory: +3 out of 4 points
You were asked to discuss one theory, including at least one proponent of the theory and four or more concepts associated with the theory. This paper considers labeling theory of deviance. The paper's treatment of labeling theory of deviance is inadequate. This paper does not discuss Howard Becker, Thomas Scheff, Thomas Szasz, and William Chambliss, who are related to this theory. This paper discusses labeling, primary deviance, and secondary deviance, which are related to this theory. This paper does not discuss . . .

Type of Crime: +1 out of 1 point
You were asked to identify one type of crime and interpret it with a single theory. This paper appears to focus on toxic pollution which is mentioned 1 times . . .

Overall Score Summary:
20 points possible
+ 2 for concepts related to the theory
+ 2 for identifying the theory
+ 2 for summarizing the theory in a paragraph
+ 3 for overall treatment of the theory
+ 1 for identifying the type of crime
+ 3 for relating the type of crime to this theory
Your score is 13.

with a program module that identifies specific kinds of reasoning and relationships among those concepts and applies the grading rubric. Each program module can be applied to a wide range of substantive topics represented by different semantic networks. For example, the modules designed to assess the deviance exercise could be used for a similar exercise regarding the family or political life. Thus, this basic essay-grading program can be extended to generate literally hundreds

or thousands of exercises combining different substantive content with different learning objectives. This should dramatically reduce further development costs for additional essay-grading modules both within sociology and in other disciplines.

Several logistical problems must be addressed for automated essay grading to be practical for a course. Students must have access to computers in order to produce papers in machine-readable form (Yang, Buckendahl, and Juszkiewicz 2001; Palmer, Williams, and Dreher 2002). Those papers must be formatted to meet certain standards. For example, hard returns at the end of each line rather than at the end of each paragraph may make it difficult for programs to recognize when paragraphs begin and end. Some system for file management is required to handle the many different student papers. A grade-recording system is needed to track student performance and report scores both to the students and to others. The University of Missouri, where this program is in use, has many computer labs for students and most students also have their own computers, so access is not a problem. The SA-Grader stores every student draft in a database along with the comments generated by the program and a summary file of scores. Instructors can review each draft and see how students have changed their papers in response to feedback. A reminder to students not to put hard returns in their texts has been sufficient to avoid formatting problems.

Potential for Deception and Cheating

A continuing concern about machine-scored essays is whether sophisticated writers can take advantage of the program's features to deceive the program into giving them a better grade than they deserve (Baron 1998; Kukich 2000; Powers et al. 2001; Rudner and Gagné 2001; Yang, Buckendahl, and Juszkiewicz 2001; Palmer, Williams, and Dreher 2002). Because the Project Essay Grade (PEG) program (Page 1966) emphasizes surface features and syntax while largely ignoring content, "a well-written essay about baking a cake could receive a high score if PEG were used to grade essays about causes of the American Civil War" (Rudner and Gagné 2001, 2). On the other hand, the Intelligent Essay Assessor (IEA) emphasizes content and largely ignores syntax. So, "conceivably, IEA could be tricked into giving a high score to an essay that was a string of relevant words with no sentence structure whatsoever" (2). (See McGee, chapter 5 in this volume, for an example of tricking the IEA scoring machine.) For this reason and others, many current applications of essay-grading programs for high-stakes assessments such as the GMAT

have every essay read by at least one human reader in addition to an assessment by a grading program (Kukich 2000).

In its current form, the Qualrus-based SAGrader program cannot discriminate between papers that are well written and those that string together key concepts. However, it does look for structures like paragraphs containing summaries, sentences linking concepts to theories, and so on. Future versions will certainly attempt to expand these capabilities to assess other aspects of writing style and rhetorical strategy. Until the program can do all these things, though, every grade assigned in the WI course by the automated essay grader will be reviewed by a human grader. In the current course, students were informed that the instructor would read every paper in addition to the machine scoring and their score from the computer was only an initial estimate of their grade. This seemed to be sufficient to encourage them to write sensible papers rather than simply string together words.

Automated Screening for Plagiarism

The program includes a built-in test for plagiarism. Each paper submitted to the automated essay grader in the WI course is compared with the database of all papers submitted for the same assignment. Papers displaying suspiciously high similarities are flagged for review by the instructor or TAs to assure that students are not plagiarizing the work of others. Of course, the system does not address all forms of plagiarism, such as copying materials from the Internet. This feature is new and has not been tested in application yet. However, we've told students the program can do this; we'd much rather prevent plagiarism than discover it.

Limitations of Scope and Depth

Perhaps the greatest concern we have about essay-grading programs is what they do *not* address. This program is able to assess important elements of student essays such as their understanding of the relationships among key concepts, their ability to use sociological concepts and perspectives to interpret their own experiences and those of others, and their ability to understand and critique empirical studies. But there are many aspects of a written paper that are not yet addressed by SAGrader or other programs. It seems likely these programs will continue to become more sophisticated and to broaden the scope of issues they examine. So far, though, SAGrader has proved to be very flexible, and we have been able to create writing exercises of considerable diversity. But there are likely to remain, at least for the foreseeable future,

important aspects of student writing that only human graders can judge.

Pilot Testing: Performance and Student Assessments

We piloted the program and the essay-grading procedures for two semesters using the deviance exercise as an extra-credit project in my section of Introductory Sociology. In a third semester we incorporated the deviance exercise, the "what is sociology" exercise, and the "evolution of community" exercise into the course as part of the required assignments. These three semesters have provided an excellent opportunity to test most of these exercises and assignments (or variations of them) and further improve them. They also permitted us to test the logistics of the process to make sure it worked smoothly.

Students were able to conveniently and easily submit their papers over the Web and receive immediate and detailed feedback. Students were asked to e-mail me if they felt they were graded unfairly; fewer than 5 percent did so. Roughly half of those were minor problems such as a phrase that was not properly recognized. Those problems were easily corrected and such problems had essentially disappeared by the third semester of pilot testing. The other half of students' complaints did not concern problems with the program. For example, one student complained that she had used other terms to indicate some of the concepts instead of the precise terms and hence she felt the program was at fault. I explained to her that part of the learning objectives of the course was to learn the appropriate technical terms.

Appended to the feedback on students' papers is a brief questionnaire in which students are asked how fairly they thought the program graded various components of the assignment and what they did and did not like about this project. In pilot tests of the deviance writing exercise, for example, students liked the essay grading, even though this was our first trial of the program and there were some imperfections in its grading. Students appreciated the immediate feedback (92 percent liked it, with 60 percent liking it a lot), the opportunity to revise their paper (92 percent liked it, with 66 percent liking it a lot), and the detailed comments (88 percent liked this aspect, with 43 percent liking it a lot). Most (65 percent) thought the initial grading was fair; 35 percent disliked the initial grading. They preferred this form of evaluation over multiple-choice tests by almost 2:1, with 47 percent preferring automatically graded essays, 26 percent preferring multiple-choice tests, and 26 percent undecided.

INTEGRATING AUTOMATED ESSAY GRADING INTO A WRITING-INTENSIVE SOCIOLOGY COURSE: TOWNSEND'S PERSPECTIVE

As director of the University of Missouri's Campus Writing Program, one of my responsibilities is to facilitate communication between the Campus Writing Board, which certifies writing-intensive (WI) designations, and discipline-based faculty, whose courses are needed from across the curriculum to satisfy the university's two-course WI graduation requirement. My writing program colleagues and I are charged with helping faculty develop academically rigorous WI courses that meet the board's criteria. In this particular case, Brent volunteered the course, but the board, faced with certifying its first machine-graded WI course, balked. Among the concerns raised were whether the machine scoring would be accurate, fair, able to provide high-quality feedback that leads to substantive revision and, not least, what "messages" would be sent to students about academic writing. In this section, I describe the process of finding common ground between the instructor and the board, a process that involved articulating skepticism diplomatically, broadening understanding on both "sides," learning new technologies, and fostering experimentation.

Writing program staff members don't recall exactly when we became aware of Professor Ed Brent's work with machine-assisted grading of writing. But the grapevine on our campus—where writing is reasonably well attended to for a large research university, in both WI and non-WI courses—had brought us news that he was up to something out of the ordinary. None of us had met him, though, on any of the committees that typically draw faculty who are interested in pedagogy and/or writing. He hadn't taught any WI courses. And although he had attended one of the faculty writing workshops we offer twice a year, it was way back in January 1987. So when he called us to inquire about the process for having his large introductory class designated as WI, we were mildly surprised and, I must admit, skeptical and even a little put off. How was it, we wondered, that a faculty member who, to our knowledge, hadn't shown recent interest in student writing could want WI status for his course—and not just any course, but one that typically enrolls 250 students? Jo Ann Vogt, our liaison to MU's classes in the social sciences and to whom I passed along this information, reacted with incredulity. "Let me get this straight. A professor wants to offer a WI class, but doesn't want to engage with the students' writing himself? Wants a machine to do the work for him? Isn't there something odd about this?"

Ed was prepared for our skepticism, though. He described his several-years'-old experiment with machine-scored writing and said he believed his project was far enough along to try out in the WI setting. No doubt discerning hesitation in my voice when I explained the WI proposal process, he offered to come to my office to demonstrate his program. I accepted, even as I wondered what the writing program was getting into. The program has a proud history of opposing standardized writing assessment. In the early 1990s, I had chaired our campus's Assessment Task Force whose main focus, it seemed, was educating faculty and administration about the drawbacks of standardized assessment of many kinds. We actively resisted a statewide impetus to assess general education (including writing) with an "off the shelf' instrument. I still see the task force's most significant achievement as having persuaded our chancellor to seek the Board of Curators' rescission of their mandate that MU students take an expensive and ineffective standardized test of general education. The curators did indeed rescind the mandate, and MU has engaged in a more responsible form of general-education assessment ever since. Additionally, for the twenty years that our WI requirement has been in place, we've successfully avoided one-size-fits-all tests of writing. So, to find myself discussing a possible WI course that would feature machine-scored writing was unexpected, to say the least.

February 18, 2004

Laptop in hand, Ed arrives at my office at the appointed hour. Our opening hellos are friendly and comfortable since, despite our not having worked together at MU, we know one another through our significant others, who both work at the local high school. We sit down, he more confident than I (in my perception) because he knows what he's going to demonstrate and I'm still skeptical, though by this time I'm also more curious than before. In lay language, Ed gives me a quick background on how the system works; still curious, I begin to wonder whether I'll follow what seems to me a technical explanation beyond my ken. "Statistical versus rule-based approaches," "parameters of regression equations," "substantive knowledge expressed through a semantic network linking key concepts," "rudimentary natural-language processing." I recognize the words, but can't think fast enough to comprehend them in the new and unfamiliar context. I flash back to David Bartholomae's concept of students inventing the university (1985) and wonder if I can invent enough leaderly acumen to maintain credibility with Ed, a senior

colleague who's been a full professor for longer than I've had my Ph.D. Realizing that a concrete example is called for, Ed opens the program on his computer and shows me a sample writing assignment, a two-page sample student response to it, and then a sample of the feedback his Qualrus system provides to the student. I follow along, though still unsure about formulating intelligent questions. He continues with an explanation of the array of responses the system can provide, based on the range of text students might enter.

Finally, something clicks and I comment, "But this assignment and the student's short response involve mainly straightforward reading and recall cognition. This isn't the in-depth critical-thinking writing that WI courses call for."

And this is when our breakthrough, of sorts, occurs. "Well, no," Ed replies. "These are exercises students do to help them acquire familiarity with the founders of sociology, the historical contexts within which they worked, their key concepts, the theories that dominate the field, and so on." Tightly structured questions that require focused responses, he points out, allow students to "rehearse" what they're learning. And if their responses don't conform to the narrow prompt, the computer tells them what's missing, and they can add to their responses to improve their scores, scores that comprise only a minor portion of their overall grade. "I'm not looking for deeply analytic thought, nor do I care about grammar and spelling with these exercises. With writing, students can assimilate ideas that simply reading or even reading with multiple-choice quizzes can't accomplish. But with 250 students each semester, machine feedback is the only way I can do it."

"Writing-to-learn," I say. "You're machine scoring revised microthemes to promote learning." Now it's Ed's turn to process my discipline's discourse. I describe the writing-across-the-curriculum pedagogies he has unknowingly adopted: short writing assignments focused on specific problems, attention to concepts over mechanics at the early stages of the process, rewriting to clarify one's ideas (e.g., Bean 1996). I am tempted to cite some of the seminal literature (Britton et al. 1975; Emig 1977) and a few of the movement's founder-practitioners (Fassler [Walvoord] 1978; Bazerman 1981; Maimon 1981; Fulwiler 1984), but I refrain so as not to appear overly eager. As we engage in further exploration of one another's work, I learn that he uses four of these short exercises to help prepare students to write a longer paper requiring synthesis and application of sociological content, and that in addition to the machine scoring, both short and long papers are read, discussed, and graded by Ed

and two graduate teaching assistants in twenty-five-student once-a-week discussion sections that accompany the twice-a-week lectures.

Before long, I realize that Ed isn't an educational charlatan using machines to do the work that he doesn't want to, as we imagined might be the case. He's a serious educational researcher whose two-decade research agenda has focused first on social interaction and later on developing computing technologies to practice research and train others to reason sociologically. He's actively working toward a future in which the two will converge, and he's anticipating the implications for both research and teaching. More relevant to the writing program's purposes, he's closer than we imagined to offering the kinds of writing and learning experiences that WI courses encourage. I ask if he'd be willing to repeat the demonstration for the writing board at its next meeting.

March 18, 2004

The Campus Writing Board, having a year earlier tabled a previous WI proposal from Ed based largely on skepticism about the machine-grading component, convenes to see his presentation. In between this meeting and his earlier demonstration for me, Ed and I have had our proposal accepted for this very chapter in Ericsson and Haswell's book; knowing this, board members listen with keen awareness of the stakes involved. He acknowledges the hesitancy they bring and the controversy that machine scoring engenders, but points out that with only two TAs for a class of 250 students, it isn't possible to assign and respond to writing in a timely enough way for students to benefit, nor can he assure that TA responses are consistent. He explains that he has developed this system because he wasn't satisfied with students' learning when he used objective tests and that machine scoring is one attempt to resolve this dilemma.

Board members observe how students enter short papers via WebCT, how the scripts Ed has written for that assignment review the text to identify required concepts, and how quickly students receive detailed feedback—usually in one or two seconds. He explains that students are invited to consult with him or their TA whether or not they have questions about the machine score (which is always tentative and subject to TA review for accuracy). After revision, students can resubmit the paper for additional machine scoring before it goes to the TA for a final score. The student's grade is the weighted average of the machine-scored draft (1/3) and the subsequent TA-scored draft (2/3). Together, the four papers account for 30 percent of the course grade. One longer paper

accounts for 45 percent. Writing, in other words, accounts for 75 percent of the total course grade. Participation in the discussion sections is 25 percent. There are no exams and no quizzes.

Questions ensue. How do the "scripts" work? How structured must the assignments be? Can assignments involve problem solving? Can students subvert the system with content-free responses? Given the tight structure of the assignments, does the program check for plagiarism? What about false positives for plagiarism? What do students think of machine scoring? How long does it take to set up a new assignment and the scripts that respond to it? Why did you rule out the existing software and design your own? Can you envision a program that discerns whether students understand hierarchies of relationships among related ideas? Could you use your system in upper-division courses as well as the introductory course?

Ed thoughtfully answers each question in turn, acknowledging the limits of the technology and its use in the classroom. Scripts contain key words, qualities associated with them, and certain patterns of argument; the computer looks for these words and patterns in the same sentence. Eventually, Ed hopes to develop scripts with the ability to identify causal and functional explanations. Assignments must be highly structured to be machine graded. Not all writing assignments should be structured this way, he says, but certainly some can be. These work to tell him whether students are learning something about sociology. Eventually, Ed hopes to develop scripts that provide feedback on transitions, paragraph length, and so on. Yes, problem-based assignments are possible. Ed tries to start with basic sociological principles and move gradually toward more intellectually challenging topics like gay marriage, for example, which have no "right answer." Yes, students could fool the system with content-free prose, but they know that Ed and the TAs skim the papers in any case; so far, no students have tried it. Yes, the program does check for suspicious similarity among papers. Given the tight structure of the assignments, there is some uniformity to them; but students are "amazingly creative" in applying different theories and organizing content differently. The questions have enough room for students to use material from a given chapter in a variety of ways. Students' evaluations show that they don't trust machine scoring as much as they trust Ed and the TAs. There's a tension, he says, between students thinking, "I got a low score, so the program is worthless" versus "I got a low score, so I better make some changes." He doesn't know how that will work out, but says it could be a problem.

Board members also want to know long it takes to set up new scripts. Ed says that now that the system is worked out, he can add new code and comments quickly. The scripts are fairly general so most of the effort goes in to changing the content, not the form, of the script. Ed ruled out existing software because most employ statistical approaches to automate the grading model; they work well for standardized writing assessment in which "mechanics" are scored, but they don't work well for his purpose, which is scoring substantive, discipline-based knowledge. Eventually—if there are certain phrases that would indicate hierarchy of ideas—Ed could code for them, but there will never be a program that looks for everything. However, many concepts in sociology are standard enough that the system works fairly well now. It wouldn't be adequate for the humanities; it might work in the physical sciences. Finally, Ed says, he couldn't claim that the program would work well in upper-division courses. He'd have to consider the course objectives and, while the present version might hold some promise at that level, it's still evolving. Over time, with a given course, it can become more useful, but no machine-scoring system will ever do everything.

Board members also want to know how the scripts are developed; who does the work? In his course, Ed developed them himself. Other content-area instructors using SAGrader could develop their own or use concept maps developed by other expert authors. He is working with publishers to make versions available in other disciplines. Board members wonder how much time is saved if all the essays are also read by human graders. Since WI courses require students to submit multiple drafts, Ed points out that he will use the program to grade first drafts of the writing exercises, and he and the TAs will grade the final drafts. The program should reduce their grading time by about half. More important, because the program gives students immediate feedback and permits them to revise and resubmit papers several times, students can submit as many as five or six versions, something they could not do with human feedback. Finally, a board member from education asks how this kind of machine scoring might translate to K–12 settings. Ed explains that since the program scores essays based on substantive content as expressed in the semantic diagrams, as long as diagrams express knowledge taught in K–12 settings, the program should work. In some cases, slightly more simplified versions of the semantic diagrams could express content appropriate for a wide range of educational levels. In other cases the semantic diagrams need not change at all; a simpler statement of the assignment with expectations appropriate for each grade level could make the program appropriate for K–12 classes.

By the end of the demonstration, Campus Writing Board members and program staff are convinced that the manner in which Ed uses machine scoring, combined with the overall design of Sociology 1000, not only does not violate the WI guidelines established in 1985, but in fact it addresses them in new and innovative ways. They vote unanimously to designate the course WI for 2004–5 and ask that Ed submit an assessment at the end of the year. Shortly after the presentation, Ed e-mails to thank us to arranging it. "I appreciate the questions and comments I got from the board members. They raised legitimate concerns, and I hope they can see that I share them. I believe the only way this program can be effective is as part of a complete course structure that provides the kinds of checks and balances needed to assure quality."

December 10, 2004

At the end of fall semester, with the first machine-graded version of a University of Missouri WI course completed, Ed reports that it "went well—but not perfectly." Some students continue to focus on format—single or double space? font and margin size?—not yet understanding that the program doesn't even look at these; sociological concepts are the primary learning goal. Some students are having difficulty submitting papers to the Qualrus server. Others are irritated by the machine scoring's imperfections, for example, not recognizing an unusual concept and awarding fewer points than deserved or not recognizing concepts that are used incorrectly and awarding more points than deserved. "Oddly enough," Ed comments on the class listserv, "few students complain about the latter."

What turns out to be the most troublesome aspect for Ed and his TAs is machine grading drafts of the longer, more complex paper. It comprises three parts: (1) identifying an important technological problem that influences work in America; (2) proposing a solution for it; and (3) designing a research study to assess the impact of the solution. Because Qualrus looks only at the whole rather than at individual parts, machine feedback is compromised. Ed notes, "We weren't happy with the program's performance on test drafts, so we graded the first draft of the term paper by hand. We will continue revising and improving the program so that it can be used more effectively for the first draft next semester."

May 26, 2005

Things go more smoothly the second time around. Ed and the TAs modify some of the writing assignments to incorporate more content from

the textbook they are using, and the Internet connection to the server is more reliable, producing less stress for students who submit on the last day. A few students, however, do not understand that the final essays are graded by course staff and are more intent on trying to fool the program than improve their papers. Ed will reduce this tendency next time by having final drafts submitted through the server just as the first draft was, so he and the TAs can call up the paper, view the program's grade and comments, and make necessary changes in the final grade. In addition to soliciting students' reactions about the grading system's fairness, Ed is adding a check item for students to indicate if they want to appeal the program's score and a text field where they can specify what they believe the program did wrong. This will provide an ongoing mechanism for quality improvement and help isolate remaining weaknesses in the program. Ed reports that, on comparing many of the first and last drafts submitted to the program, "it's encouraging to see that they often improve substantially." Students continue to like the immediate and detailed feedback, he says, as well as the opportunity to revise their papers. At this point in the experiment, Ed and his TAs believe the system offers a sensible way to offer students writing opportunities that replace multiple-choice tests in large-enrollment classes.

As we submit our chapter, the Campus Writing Board still awaits the results of this first year's trial with machine-scored writing at our university. A specially convened summer board meeting will determine whether machine-scored WI versions of Sociology 1000 will be offered in 2005–6. At this same time, however, other questions also loom for the field of composition studies. We see that recent policy statements that attempt to shape good practice in writing assessment and machine scoring may not have fully anticipated the pedagogical applications of technology. Ed's work problematizes these new policies. For example, the section on electronic rating of placement tests that is part of the Conference on College Composition and Communication's "Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments" (2005) states unequivocally that "all writing should have human readers, regardless of the purpose of the writing" (789). This section also claims that (1) "writing to a machine . . . sends a message [that] writing . . . is not valued as human communication"; (2) "we can not know the criteria by which the computer scores the writing"; and (3) "if college writing becomes *to any degree* [emphasis added] machine-scored, high schools will begin to prepare their students to write for machines." The overall statement ends by noting that machine scoring is being considered for use in writing

centers and for exit tests, and its unambiguous conclusion is, "We oppose the use of machine-scored writing in the assessment of writing."

The CCCC "Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments" seems not to anticipate classroom use of the kind to which Ed Brent is applying machine scoring. In his class, machine scoring is a complement to human reading, students do know the criteria by which the computer arrives at their feedback, and if high school writing teachers did inculcate the advantages of writing-to-learn and prepared students to respond to content-driven microthemes, students would likely benefit. In its strident "regardless-of-the-purpose" stance, the digital position statement does not acknowledge that—*when used responsibly* (as I would argue Ed is doing in Sociology 1000) *and when not used as the sole or even primary determiner of grades* (as Ed is not doing)—machine-scored writing might assist and enhance learning, as is its purpose in his large-enrollment course. Ironically, the other choice available for Sociology 1000 is scantron-graded multiple-choice tests, a machine-scored form of assessment that does not enhance learning. Given the impediment of responding to writing in a class of 250 students with three instructors, using technology to assist learning rather than test objective knowledge seems the preferred alternative.

The earlier CCCC "Writing Assessment: A Position Statement" (1995) lays out the profession's best thinking on ways to "explain writing assessment to colleagues and administrators and secure the best assessment options for students" (430). Few would disagree with this statement's cautions against high-stakes, standardized assessment of writing. But many would probably be surprised by the number of positive correlations between the statement's recommendations and Ed Brent's use of machine-scored writing in Sociology 1000. Using the language of the statement, a partial list includes: providing assistance to students; its primary purpose governs its design and implementation; students clearly understand its purpose (learning objectives appear on each assignment); it elicits a variety of pieces over a period of time; it is social (students freely discuss their machine-scored writing experiences online and in discussion sections); reading is socially contextualized (reading the course material is necessary for the machine-scored writing); a variety of skills in a diversity of contexts is employed (different genres, audiences, occasions, and readers are involved); the assessment is used primarily as a means of improving learning; it does not focus on grammatical correctness and stylistic choice and does not give students the impression that "good" writing is "correct" writing; large amounts

of institutional resources were not used to design or implement the machine scoring; students are encouraged to plan, draft, and rewrite; students write on prompts developed from the curriculum that are grounded in "real-world" practice; students know the purpose of the assessment, how the results will be used, and how to appeal a score; the faculty member played a key role in the design of the assessment; the faculty member participates in reading and evaluating student writing; the faculty member assures that the assessment supports what is taught in the classroom; and the faculty member continues to conduct research on writing assessment, particularly as it is used to help students learn and to understand what they have achieved. This is a long list of positive correlations between composition studies' professional recommendations and the program Ed has designed and is using.

In light of the challenges that Ed's example offers to the thinking in composition studies to date, it is time for composition specialists to revisit our professional policies and practices. Such revisiting is to be expected, given the changes that technology has wrought in the teaching of writing, not just in the past couple of decades but over the centuries.

CONCLUSION

Ed began this chapter by pointing to the promise of automated grading programs for writing. Indeed, impressive claims can be made for them. Pilot tests of SA-Grader suggest that it can reduce costs for large classes, provide immediate and detailed feedback in a manner students appreciate, and apply to a wide range of exercises addressing substantive concepts and theoretical perspectives. However, there are serious issues to be considered if automated grading is to be used both appropriately and successfully. Ed and I both believe there is a place for machine-scored writing, so long as the concerns we raise are carefully considered. The role of machine-scored writing will likely always be limited, but that role will surely evolve as technologies mature. We don't believe that essay-grading programs will ever become a panacea for writing classes, nor that they should or will replace human teachers. But *when used responsibly* they can make writing assignments such as the writing-to-learn exercises we described above feasible in a wider range of courses. And, when incorporated into courses in ways that minimize their weaknesses, they can provide a meaningful enhancement to student performance.