15

WHY LESS IS NOT MORE

What We Lose by Letting a Computer Score Writing Samples

William Condon

Earlier in this volume, Rich Haswell (chapter 4) questions the validity of machine scoring by tying it to holistic scoring methodologies—and I certainly agree with his critique of timed writings, holistically scored. However, I want to suggest that questions about machine scoring differ from questions about holistic readings. Machine scoring involves looking back even farther in the history of writing assessment—to indirect testing. Several essays in this collection describe the basic methodology behind machine scoring. The computer is programmed to recognize linguistic features of a text that correlate highly with score levels previously assigned a text by human raters. Thus "trained," the machine searches essays for those features, assigning them the score levels indicated by anywhere from thirty to fifty linguistic markers. In this way, the machine achieves as much as 98 percent agreement with human raters (which, of course, means that it is no better than 2 percent *less* reliable than human raters).

What we want to notice here is that the machine does not in any sense *read* a text. It simply searches for a feature (periodic sentences, conjunctive adverbs, topic-specific vocabulary, vocabulary or concept mapping, etc.) and assigns a score based on how many of those features it finds and how frequently it finds them. This is not really comparable to holistic scoring, since the machine is incapable of forming an overall impression of an essay—or, for that matter, any kind of *impression* about anything. As I have written elsewhere (Condon and Butler 1997), the machine is incapable of understanding the difference in meaning between these two sentences:

> The roast is ready to eat.
> The tiger is ready to eat. (2)

In other words, nothing in the machine's scoring process takes into account the content, the semantic effectiveness, or the rhetorical choices in the essay being scored. Instead, the machine looks at physical

features of a text that, separately, are *associated* with a certain level of performance. This process more closely resembles the old multiple-choice question tests, which purported to judge writing proficiency by asking a set of questions that focused on a range of abilities associated with good writing: vocabulary, syntactic knowledge, error recognition, and the like. In other words, instead of a step forward, or even marking time, machine scoring represents a step backward, into an era when writing proficiency was determined by indirect tests.

For the moment, though, and just for the sake of argument, let us assume that machine scoring lives up to the representations of its various promoters. We can come back later in this essay to consider the more worthwhile assessment alternatives that machine-scored timed samples, because cheaply administered, threaten to displace. For now, let us *pretend* that the machine reads as human raters do, that it is capable of making fine judgments about writing ability as a whole construct, that its scores are just as good as those rendered by human raters. In fact, arguing over the claims advanced by the testing agencies may engage us in chasing after red herrings, since the real question is not whether machines can do what the agencies claim, but whether machine-scored timed samples are better than the alternatives—or at least whether a cost-benefit analysis would come out in favor of machine scoring. So we need to look first at the losses—the testing agencies have been quick to point out the gains—involved in using computers to score timed writings, and then, later in this essay, we need to consider the alternatives to stepping backward to indirect assessments of writing.

If we grant, *arguendo*, all the claims in favor of machine scoring as being similar to what human raters do, where does that leave us? If the machine can score as accurately as—and more efficiently than—human raters, that represents a gain. But what are the corresponding losses? We need to examine that list before we decide that machine scoring—even if it could be as good as advertised—should take the place of human raters.

First, and perhaps most basically, in any test type that is administered nationally, rather than locally, we lose control over the construct: *writing*. What we assess is dictated to us by an outside agency—and specifically, in this case, by the capacity of the rating machine. Second, samples must be short, thus preventing the writer from taking an original approach to a topic, coming up with a different approach, organizational pattern, or even vocabulary items, which would inhibit the machine's ability to fit the sample within its set of algorithms. So the writing sample is frequently limited to what a writer can produce in twenty minutes or half an hour.

Given the constraint of time, topics must be far simpler than the topics most teachers would use in class, even at the beginning of a term. Both these factors limit the face validity of the sample. In addition, fast, off-the-cuff writing typically cannot contain much depth or complexity of thinking; practically speaking, the writer simply has no time to do more than sprint to finish an essay that is on topic—and that is so short it hardly deserves to be called an essay. Such a sample, however it is scored, cannot tell us much about a student's writing ability, because the sample's validity is so narrow that it cannot test very much of the construct.

The limitations on face validity mean that we can draw only very limited conclusions from the sample. We can, for example, as Edward White (1994) has pointed out, tell whether the test taker can produce competently formed sentences. We can make some conclusions about the fluency of the writing and about the writer's ability, unassisted, to produce more or less mechanically correct prose. In other words, we can make the kinds of judgment that might allow us to place a writer into a very basic course that deals with sentence-level problems or a higher one that might begin with writing paragraphs or short simple essays. Such an assessment is not useful to most four-year colleges, which are typically prevented from offering such basic courses by legislatures that insist that four-year colleges and universities offer only "college-level" courses. In sum, then, four-year colleges lose the ability to make any sort of useful distinction or ranking among their entering students, since (1) all should be performing above the level that such a short timed machine-scored sample can measure; and (2) even if students perform at a lower level, the four-year school can offer no course to help those students. And the same problem faces the two-year college for all students who are ready for college-level writing: since the sample measures a construct that is significantly below what college-level courses offer, the institution can have little confidence in a decision that places a student into college composition or even into the foundational course immediately below college composition. Of course, community colleges are generally able to offer courses at a sufficiently low level that this distinction might apply—yet the number of such students is quite low, so the question of economics returns. But more to the point, sorting students by ability is supposed to result in classes where the range of ability varies but is manageable. Such short, limited samples cannot provide enough information to make such finer judgments. This latitude leaves teachers holding the bag, in classes where the range of students' abilities is potentially so broad as to make teaching more difficult than it needs to be—or should be.

If the machine could score longer, more topically complex samples, that would be an improvement, but really the objections above extend, to some degree, to even the best timed writing tests, scored by human raters. My own program offers students a choice of topics, all of which solicit more than one sample—one analytic or argumentative and another reflective—that respond to a short selection of text, and the testing session allows two hours for completion. Yet even this far more robust sample, as Diane Kelly-Riley and I have demonstrated elsewhere (2004), is not sufficient for test takers to incorporate critical thinking into their samples—at least not without cost. When we score a set of timed writings for placement and then again for critical thinking, the resulting scores actually show a negative correlation. In effect, if students choose to think, their placement scores suffer (see White 1994 for confirmation of this phenomenon). For several reasons, this negative correlation is not surprising, but it means that if we want to assess students' readiness for college writing—or, beyond that, whether they should be exempt from the course—even a much more robust sample than the computer can score is incapable of reaching the competencies involved in such a decision. Indeed, as I will discuss later in this essay, timed writing samples are themselves of such limited validity that their ability to provide this kind of information is low. The overwhelming majority of students simply cannot produce such evidence in such a limited sample of writing. To the extent that the placement decision depends on anything but the most basic aspects of good writing, these samples also lack predictive validity.

Losing control of the construct involves a second loss: the assessment inevitably takes place on a national level, rather than on the local level. Writing prompts are designed by experts at national, even international testing firms. Such prompts almost certainly have little to do with local curricula, and they may well be inappropriate for a local student population. In Washington State University's entry assessment, the prompt asks students to respond to a short argumentative reading because the ability to analyze and interpret a text, as well as join in conversation with the text, is central to the curriculum of English 101. Similarly, we demand that students summarize the author's position as context for the student's own position on the issue. This demand parallels the 101 curriculum, and it responds to an assignment type that many teachers of first-year students tell us they use, whether in English 101 or other courses across the curriculum. In the consideration of the diminished construct, we saw what we lose in decision-making ability. By using

someone else's topics, we see what we lose in our ability to be sure that the instrument actually measures a construct that is relevant to local curriculum and expectations. In other words, by importing topics and judgments that are national, rather than local, we lose important aspects of systemic validity.

Of course, using a large-scale test that is national in scope means that the criteria used in judging the sample probably do not match local expectations either. While the testing services offer customized samples, these are far more expensive. The default test uses the testing agency's topic and the testing agency's raters to set the machine's parameters, so there is no relationship between the test's results and local curriculum, local standards, or local course sequences. In other words, local administrators are little better off than when they set a cutoff score in (mis)using the SAT verbal or the ACT English score for placement: beginning with a "best guess," the program administrator adjusts the cutoff, over time, until placements seem roughly to fit course levels. Such a procedure never results in placements that are as accurate as possible, and such a process mistreats several terms' worth of students, until the level stabilizes. The best assessments are local, since in the local context teacher/raters understand their curriculum and their expectations, so that they can make firmer judgments matching a particular sample to a particular level of instruction with which those teacher/raters have firsthand experience. Teacher/raters who actually teach in the program they place students into know what the beginning writing of successful students looks like, and they can make placements according to an "expert rater" system. The advantages of such local assessments have long been documented (see Haswell 2001 and Smith 1993 for examples at different institutions), so we should be reluctant to give up these benefits.

A third, and perhaps even more costly loss occurs when the machine stands in as one rater of two (one human score paired with one machine score, with disagreements resolved by a second human rater). In assessments that use two (or three) human raters, conversations about writing, about writing standards, about judgments of quality occur. In addition, when local assessments use writing teachers as raters, those teachers share a great deal of lore about course expectations, signs of student ability, curriculum, and so on. Teacher/raters bring their knowledge of the instructional context with them, and that knowledge aids in making more accurate decisions. During the assessment, they learn a great deal about incoming students, information that helps them as they move back into the classroom. This system is reiterative

and cumulative, constantly feeding a rich knowledge set from instruction into assessment and from assessment back into instruction. These conversations serve a number of other purposes as well. Teacher/raters define the construct operationally for themselves, and they carry that common sense of the construct into planning their own course assignments and activities.

These sessions also serve as powerful faculty development. Teachers talk not only about quality but also about strategies: how might we handle this student in this course? Why should we realize that this student probably could not succeed at the assignments we typically offer in a particular course, while he or she might well succeed at the tasks offered in another? What sorts of assignment might result in more of these writers succeeding in our course?

Local assessments, which typically employ teachers as raters, produce more valuable and more interesting outcomes than merely a score with which to establish a ranking upon which a placement can be based. Move the assessment away from the instructional context *and* plug a machine in as one rater, and we break the cycle. Those interactions simply cannot happen. Taking the assessment out of its context drastically reduces the information available from the assessment. While we might argue that even a poor assessment, done locally, produces benefits that make it worth the trouble and expense, clearly we could not make such an argument in favor of machine-scored timed writings. If the scores themselves are not worth the expense and trouble, then the test also is not—because the scores are all we get from such an assessment.

Fourth, in various ways aside from those discussed above, a timed machine-scored sample takes away local agency. The shorter the sample, the lower the level of confidence that students and teachers have in it. Indeed, in my own experience working in four universities' assessment programs and consulting with dozens more, if the sample requires less than an hour to complete, teachers routinely administer a second sample on the first day of class in order to make a second judgment about whether a given student belongs in the course. This wastes time and effort, of course, but the point here is that if a more robust, human-scored sample is below the teachers' trust threshold, then all machine-scored timed samples are necessarily below this trust threshold, and so will create duplication of effort. In addition, students distrust and resent timed samples, even the ones that offer extended times and multiple topics and genres. Their level of confidence and investment being low, their performance may well suffer, but the main problem is that they

begin the course resenting the assessment process and often convinced that they do not belong there, particularly if the course is at a lower level than they had hoped for or at a higher level than they had expected. The fact that the topic does not match the course's curriculum also saps agency from the teachers, who have been told, implicitly, that they are not qualified to make these judgments, that their institution does not trust them to make those judgments, or that their institution does not care enough about the students to pay the teachers to make those judgments. Any or all of these messages create an unhealthy level of cynicism and a sense of powerlessness among the teacher corps. No matter the economic benefit—even if machine scoring were free—these costs outweigh the advantages.

What we see in this cost-benefit analysis is that machine scoring's principal advantage—economy—comes at too great a cost. Institutions are tempted to adopt machine scoring because the cost of assessment is borne by the student, and that cost (most testing firms charge between $4 and $8 per sample) is lower than the cost of operating a local assessment (indeed, the institution's cost goes to zero). At my own institution, students pay a $12 fee for the Writing Placement Exam. In return, they sit for two hours and write two samples that are tailored to our English 101 curriculum. Such a test has higher face and systemic validity than a single sample written in only one-fourth of the time could possibly yield. Our students also move into classes in which instructors are better prepared, because the teachers know a great deal more about what students can do, what tasks they are ready for, where their zone of proximal development is. Thus, the higher fee comes with much higher value. Even if the institution must bear the cost—many are not allowed to charge separate fees for such an assessment—the payoff in faculty development alone seems worth the price and worth the trouble of offering a local assessment and using local faculty as raters. Machine scoring simply cannot compete economically, as long as we consider *all* the costs of employing it.

Aside from this cost-benefit analysis, another issue looms large: assessment has moved ahead since the advent of the timed writing sample in the late 1960s. Today, the demand for outcomes-based assessments that respond to benchmarked competencies drastically reduces the usefulness of any timed sample. For this and any number of other reasons, we can and should do better than timed writing tests, no matter how they are scored. Over the past two decades, since Belanoff and Elbow's (1986) landmark article on a system of programwide portfolio-based writing

assessment, the field of writing assessment has developed a robust set of tools, from portfolios to various other kinds of performance assessment based on actual student learning outcomes.

Less robust forms of assessment entail losses. Indirect tests are context-free: they do not connect with a student's curriculum, nor do they take into account the learning that goes on in a given classroom. Direct tests are context-poor. While they are based on an actual sample of a student's writing, they are so tightly controlled—in topic, time for writing, genre, and so on—that they provide only the merest glimpse into a person's overall writing competencies. Various forms of performance assessment, in contrast, are context-rich (Hamp-Lyons and Condon 1993). They not only offer a far better survey of an individual's abilities, they also bring with them artifacts from the curriculum and the classroom (assignments, for example, as well as the writer's reflections on the learning process), so that we can begin to assess writing in ways that can feed back into the classroom, resulting in improved instruction. We can also use context richness to help us make judgments about where we might improve instruction, curriculum, and course design in order to boost student performance. These more robust assessments involve looking directly at the work products students create in their classes. Therefore, this class of assessment values, rather than undermines, what happens between student and teacher, between student and student. Outcomes assessment focuses directly on what students can or cannot do, and it emphasizes the importance of doing well in class, since the effort there translates directly to results on an assessment. Finally, the reverse is also true: students are clearly invested in earning a high grade in a course, so we need not question their effort on course assignments (or if we do, at least we can say that such a level of effort is typical of a given student). We know, on an outcomes-based performance assessment, that we are getting the best effort a student will give. The same is simply not true of timed writing samples.

Since the essays that computers are able to score must be short and tightly controlled by topic (else the correlations will be too low to produce a reliable score), the result is an even more limited sample than is collected in the usual direct test, holistically scored. Such a limited sample can provide a very rough—and not very fair—ranking of writing samples (note: not of writers by their abilities). This ranking tells a teacher almost nothing about a student's performance, so it provides no feedback into the writing classroom, no information that either the teacher or the student can use to improve. As Brent and Townsend (in

chapter 13 of this volume) indicate, although there may be some tasks (i.e., short-answer exams, brief response papers) that may fall within the scope of the construct reached by machine-scored timed writings, these classroom tasks are typically not central to judging student performance there—and none require responses to topics of which students have no knowledge or for which students have had no chance to prepare. We should only use assessments for placement that for the most part address the kinds of task students will face in the classrooms for which they are headed. And any exit assessment or outcomes-based evaluation should of course depend primarily on work products central to evaluating whether students have achieved the expectations placed upon them in the course. Again, outcomes-based performance assessments address what teachers have actually asked students to learn, and these assessments provide information about whether teachers are asking students to address all they should be.

Portfolio-based writing assessments provide a clear example of these benefits, and since these assessments have been conducted successfully for almost two decades, they provide a realistic alternative for both larger- and smaller-scale assessments. Conversations around portfolios are rich and rewarding, again resulting in improved instruction both for individual teachers and across writing programs (Condon and Hamp-Lyons 1994). Performance assessments generally—and portfolios specifically—promote conversation about learning. As students assemble portfolios, they consult with their teachers and their peers. As teachers read and rate portfolios, they consult each other during norming sessions and, typically, while evaluating the difficult cases (Leonhardy and Condon 2001; Condon and Hamp-Lyons 1994). No automated-scoring program can assess a portfolio: the samples are too long, the topics often differ widely, and student writers have had time to think, to work up original approaches, and to explore source materials that help promote more complex thinking. Still, even if computers *could* make such judgments, these valuable conversations simply could not take place, so that the assessment process would exclude the one aspect that teachers—whether writing teachers or not—regularly report as the most valuable form of faculty development they have access to. (Compare Belanoff and Elbow 1986; and, to demonstrate that even ETS knows the value of these conversations, see Sheingold, Heller, and Paulukonis 1995; Sheingold et al. 1997.)

Beyond programmatic benefits, portfolios incorporate data that enable evaluation on the institutional level. Performance assessments

provide artifacts that speak to the whole process of learning to write. We see multiple samples, produced under normal writing conditions. We can see assignments, syllabi, reflections about the learning process. These artifacts provide data for accreditation purposes, for far more robust accountability measures, and for a program's or an institution's internal evaluation processes. Such high-stakes assessments of student learning outcomes can be mounted separately from placement tests, exit assessments, and program evaluations, of course, but the most sensible and economical assessments take account of each other so that data from one can provide benchmarks for the next and so that, taken together, these assessments provide a look at a student's whole educational experience along a given dimension (e.g., writing, critical thinking, quantitative reasoning). Indirect tests of any kind—or even timed direct tests, whether scored by humans or by machines—provide none of these data.

These and other losses suggest that machine scoring takes us in several wrong directions. At the very moment when state and national legislatures and accrediting agencies are demanding greater accountability—and basing that accountability on student learning outcomes—the machine-scoring process robs us of the ability to provide the fuller and more complete information about students' learning and about their achievements. At the very moment when performance assessments are helping promote consistency in writing instruction across classrooms, machine scoring takes us back to a form of assessment that simply does not reach into the classroom. At the very moment when better, more valid, more thoughtful, more accessible forms of assessment have made assessment the teacher's friend, machine scoring promises to take us back to a time when assessment was nothing but a big stick for beating up on teachers. At the very moment when writing assessments have produced extremely effective engagements of assessment with instruction, machine scoring promises to take assessment back *out* of the learning process. Perhaps F. Scott Fitzgerald, in another context, has characterized the machine-scoring initiative best: "And so we beat on . . . borne back ceaselessly into the past."