4

AUTOMATONS AND AUTOMATED SCORING

Drudges, Black Boxes, and Dei Ex Machina

Richard H. Haswell

Her name really is Nancy Drew. Like her fictional namesake, she is into saving people, although more as the author of a mystery than the hero of one. She teaches English at a high school in Corpus Christi, Texas, and according to the local newspaper (Beshur 2004), she has designed a software program that will grade student essays. The purpose is to help teachers save students from the Texas Essential Knowledge and Skills test, which must be passed for grade promotion and high school diploma. Her software will also save writing teachers from excessive labor, teachers who each have around two hundred students to protect (plus their jobs). "Teachers are going to be able to do more writing assignments, because they won't have to grade until all hours of the morning," says a school director of federal programs and curriculum from Presidio, across the state—"I'm looking to earmark our funds." That will be $799 for their campus license, according to Drew, who predicts that sales will reach half a million the first year alone.

What the administrator in the Presidio school district will be getting for his $799 is not clear, of course. Drew cannot reveal the criteria of the program—trade secret—although she allows that they include "capitalization and proper grammar among other standards." Nor does she reveal any validation of the program other than a "field study" she ran with her own students, for extra credit, in which the program "accurately graded students' work." The need for the program seems validation enough. Drew explains, "There's just not time to adequately read and grade the old fashioned way. That's what is going to make this software so popular. It's user friendly and teacher friendly." She calls her program "the Triplet Ticket" (Beshur 2004).

In the capitalistic oceans of automated essay scoring, where roam Educational Testing Service's e-rater, ACT's e-Write, and the College Board's WritePlacer, the Triplet Ticket is small fry. But in research, design, and marketing, Nancy Drew's coastal venture obeys the same evolutionary drives as the giants of the open sea. Demand for the commodity rises from educational working conditions and the prior existence of huge testing mandates legislated by state and union. The design relies on algorithms approximating writing criteria that address standards already fixed in the curriculum. The exact nature of the algorithms is kept secret to protect the commodity (proprietary interests) and sometimes to protect the testing (test security). The validation of the software is so perfunctory that the product is sold before its effectiveness is known. The benefits are advertised as improving teaching and teachers' working lives, especially the hard labor of reading and responding to student essays. Yet the product is not promoted through teachers and students, although it is through everybody else, from legislators to administrators to the newspaper-reading public. No wonder Nancy Drew thinks the Triple Ticket will be a hit. Given the rapid commercial success of the giants, she might well have asked herself, how can it fail?[1]

I have a different question. Probably this is because I'm a writing teacher who feels good about the way he responds to student essays and who doesn't have any particular yen to pay someone else to do it for him, much less someone doing it through a hidden prosthesis of computer algorithms. I'm also a writing teacher who understands the rudiments of evaluation and can't imagine using a writing test with no knowledge about its validity. I'm also human and not happy when someone changes the conditions of my job without telling me. As such, I guess I speak for the majority of writing teachers. Yet here we are watching, helpless, as automatons take over our skilled labor, as mechanical drones cull and sort the students who enter our classrooms. So my question is this: how did we get here?

To answer this question I am going to set aside certain issues. I'm setting aside the possible instructional value of essay-analysis programs in providing response to student writers—both the fact that some programs are highly insightful (e.g., Henry and Roseberry, 1999; Larkey and Croft, 2003; Kaufer et al. in press) and the fact that other programs (e.g., grammar- and style-checkers) generate a sizeable chunk of feedback that is incomplete, useless, or wrong. I'm setting aside the Janus face the testing firms put on, officially insisting that automated scoring should be used only for such instructional feedback yet advertising it for placement (the name "WritePlacer" is not that subtle). I'm setting aside the fact that, no matter what the manufacturers say, institutions of learning are stampeding to use machine scores in order to place their writing students, and they are doing it with virtually no evidence of its validity for that purpose. I'm setting aside the fact that in 2003 El Paso Community College, which serves one of the most poverty-stricken regions in the United States, itself set aside $140,000 to pay the College Board for ACCUPLACER and Maps. I'm setting aside other ethical issues, for instance the Panglossian, even Rumsfeldian way promoters talk about their products, as if their computer program lies somewhere between sliced bread and the brain chip (Scott Elliot, who helped develop IntelliMetric, the platform for WritePlacer, says that it "internalizes the pooled wisdom of many expert scorers" [2003, 71]). I'm setting all this aside, but not to leave it behind. At the end, I will return to these unpleasantries.

DRUDGES

> *We love [WritePlacer] and the students think we are the smartest people in the world for doing essays like that.*
>
> —Gary Greer, Director of Academic Counseling,
> University of Houston–Downtown

I will return to the issues I've set aside because they are implicated with the history of writing teachers and automated scoring. We writing teachers are not ethically free of these unsavory facts that we would so much like to bracket. We are complicit. We are where we are because for a long time now we have been asking for it.

Not a happy thought. Appropriately, let's begin with an unhappy piece of history. From the very beginning the approach that writing instruction has taken to computer language analysis has ranged from wary to hands off. It's true that programmed-learning packages, which started to catch on in the mid-1950s, were hot items for the next twenty years, often installed in college programs with government grants: PLATO at the University of Illinois, TICCIT at Brigham Young University, COMSKL at the University of Evansville, LPILOT at Dartmouth, and so on. But teachers—not to speak of students—soon got bored with the punctuation and grammar drill and the sentence-construction games, and found a pen and a hard-copy grade book easier to use than the clunky record-keeping functions. They read in-discipline reviews of the programs insisting that the machinery was not a "threat" to their livelihood, and eventually they sent the reels and the disks and the manuals to gather dust at the writing center (Byerly 1978; Lerner 1998).

Style-analysis programs suffered a similar rejection, albeit of a more reluctant kind. At first a few enthusiastic souls wrote their own. In 1971

James Joyce—really his name—had his composition students at Berkeley compose at an IBM 360 using WYLBUR (a line editor), and he wrote a program in PL/I (a programming language) that produced a word concordance of each essay, to be used for revisions. But ten years later he was recommending teachers use UNIX programs developed at Bell Laboratories in the late 1970s, because they were ready-made and could be knitted together to generate vocabulary lists, readability formulas, and frequency counts of features of style, all on a microcomputer (Joyce 1982). The commercial side had seen the salability of style-checkers and were using their greater resources to beat the independent and unfunded academics to the mark. The year 1982 marks the threshold of the microcomputer with affordable memory chips—the most profitable vehicle for style, spelling, and grammar-checkers—and IBM and Microsoft were ready with the software to incorporate into their word-processing programs. Long forgotten were Mary Koether and Esther Coke's style-analysis FORTRAN program (1973), arguably better because it calculated word frequency and token words, Jackson Webb's WORDS (1973), which tried to measure initial, medial, and final free modification, and Robert Bishop's JOURNALISM (1974), which reported sentence-length variance—forgotten along with WYLBUR and PL/I. Many of the homegrown programs, such as the Quintilian Analysis, were arguably worse, certainly worse than slick and powerful programs such as Prentice-Hall's RightWriter, AT&T's Writer's Workbench, and Reference Software's Grammatik.[2] To this takeover the composition teachers were happy to accede, so long as they could grumble now and then that the accuracy rate of the industry computer-analysis software did not improve (Dobrin 1985; Pedersen 1989; Pennington 1993; Kohut and Gorman 1995; Vernon 2000; McGee and Ericsson 2002).

The main complaint of writing teachers, however, was not the inaccuracy of the mastery-learning and style-analysis programs but their instruction of students in surface features teachers felt were unimportant. Yet the attempts of the teachers to write less trivial software, however laudable, turned into another foray into the field and then withdrawal from it, although a more protracted one. The interactive, heuristic programs written by writing teachers were intelligent and discipline based from the beginning: Susan Wittig's Dialogue (1978), Hugh Burns and George Culp's Invention (1979), Cynthia Selfe and Billie Walstrom's Wordsworth (1979), Helen Schwartz's SEEN (Seeing Eye Elephant Network, 1982), Valerie Arms's Create (1983), William Wresch's Essay Writer (1983), to name some of the earlier ones. In 1985 Ellen McDaniel listed forty-one of them. But where are they now? Again, industry's long arm secured a few, and the rest fell prey to our profession's restless search for a better way to teach. WANDAH morphed into the HBJ Writer about the same time, the mid-1980s, that CAI (computer-assisted instruction) morphed into CMC (computer-mediated communication). In part discouraged by research findings that computer analysis did not unequivocally help students to write better, and in part responding to the discipline-old creed that production is more noble than evaluation, composition teachers and scholars switched their attention to the siren songs of e-mail, chat rooms, and hypertext. And true to the discipline-old anxiety about the mercantile, they associated a mode of instruction they deemed passé with the ways of business. In 1989 Lillian Bridwell-Bowles quotes Geoffrey Sirc: "Whenever I read articles on the efficacy of word processing or text-checkers or networks, they always evoke the sleazy air of those people who hawk Kitchen Magicians at the State Fair" (86).

The discipline's resistance to computer analysis of student writing was epitomized early in the reaction to the first attempt at bona fide essay scoring, Ellis Page and Dieter Paulus's trial, realized in 1966 and published in 1968. Wresch (1993), Huot (1996), and McAllister and White in chapter 1 of this volume describe well the way the profession immediately characterized their work as misguided, trivial, and dead end. Eighteen years later, Nancarrow et al.'s synopsis of Page and Paulus's trial holds true to that first reaction: "Too old, technologically at least, and for many in terms of composition theory as well. Uses keypunch. Concentrates on automatic evaluation of final written product, not on using the computer to help teach writing skills" (1984, 87). In the twenty years since that judgment, Educational Testing Service's Criterion has already automatically evaluated some 2 million "final written products"—namely, their Graduate Management Admission Test essays.

If today Page and Paulus's trial seems like a Cassandra we resisted unwisely, to the ears of computer insiders in 1968 it might have sounded more a Johnny-come-lately. Composition teachers had come late to the analysis of language by computer. By 1968 even scholars in the humanities had already made large strides in text analysis. Concordances, grammar parsers, machine translators, analyses of literary style and authorship attribution, and machine-readable archives and corpora had been burgeoning for two decades. Conferences on computing in the humanities had been meeting annually since 1962, and *Computers and the Humanities: A Newsletter* was launched in 1966. It was nearly two decades later that the first conference on computers and composition teaching was held (sponsored by SWRL Educational Research and Development, in Los Alamitos, California, in 1982) and their first journals appeared (*Research in Word Processing Newsletter* and *Computers and Composition* in 1983, *Computer-Assisted Composition Journal* in 1986). By then text analysis elsewhere in the humanities had already reached such exotic lands as Mishnaic Hebrew sentences, Babylonian economic documents, and troubadour poetry in Old Occitan. Between 1968 and 1988, the only articles on computer analysis of student writing in the general college composition journals stuck to grammar-checkers, style-checkers, and readability formulas. Even summaries of research into computer evaluation typically executed a perfunctory bow to Page and Paulus and then focused on style analysis, with caveats about the inability of computers to judge the "main purposes" of writing, such as audience awareness and idea development, or even to evaluate anything since they are only a "tool" (Finn 1977; Burns 1987; Reising and Stewart 1984; Carlson and Bridgeman 1986).

I pick the year 1988 because that is when Thomas Landauer says he and colleagues first conceived of the basic statistical model for latent semantic analysis, the start of a path that led to the commercial success of Intelligent Essay Assessor. It's worth retracing this path, because it follows a road not taken—not taken by compositionists. Statistically, latent semantic analysis derives word/morpheme concordances between an ideal or target text and a trial text derivative of it. It compares not individual words but maps or clusters of words. Historically, this semantic enterprise carried on earlier attempts in electronic information retrieval to go beyond mere word matching (the "general inquirer" approach), attempts at tasks such as generating indexes or summaries. In fact, latent semantic analysis's first payoff was in indexing (Deerwester et al. 1990; Foltz 1990). In 1993, it extended its capabilities to a much-studied problem of machine analysis, text coherence. The program was first "trained" with encyclopedia articles on a topic, and after calculating and storing the semantic maps of nearly three thousand words, used the information to predict the degrees of cohesion between adjoining sentences of four concocted texts. It then correlated that prediction with the comprehension of readers (Foltz, Kintsch, and Landauer 1993). A year later, latent semantic analysis was calculating the word-map similarity between a target text and students' written recall of that text and correlating the machine's estimate with the rates of expert graders (Foltz, Britt, and Perfetti 1994). By 1996, Peter Foltz was using a prototype of what he and Thomas Landauer later called Intelligent Essay Assessor to grade essays written by students in his psychology classes at New Mexico State University. In 1998, Landauer and Foltz put Intelligent Essay Assessor online after incorporating as KAT, or Knowledge Analysis Technologies. In the next few years their essay-rating services were hired by Harcourt Achieve to score General Educational Development test practice essays, by Prentice-Hall to score assignments in textbooks, by Florida Gulf Coast University to score essays written by students in a visual and performing arts general-education course, by the U.S. Department of Education to develop "auto-tutors," and by a number of the U.S. armed services to assess examinations during officer training. In 2004, KAT was acquired by Pearson Education for an undisclosed amount of money.

I dwell on the history of Intelligent Essay Assessor because it is characteristic. We would find the same pattern with e-rater, developed during the same years by Jill Burstein and others at ETS and first used publicly to score GMAT essays in 2002, or with IntelliMetric, developed by Scott Elliott at Vantage Laboratories, put online in 1998, and making its first star public appearance as the platform for College Board's WritePlacer, the essay-grading component of ACCUPLACER, in 2003. The pattern is that automated scoring of essays emerged during the 1990s out of the kinds of computer linguistic analysis and information retrieval that writing teachers had showed little interest in or had flirted with and then abandoned: machine translation, automatic summary and index generation, corpora building, vocabulary and syntax and text analysis. Researchers and teachers in other disciplines filled the gap because the gap was there, unfilled by us researchers and teachers in writing. All the kinds of software we abandoned along our way is currently alive, well, and making profits for industry in foreign-language labs and ESL and job-training labs, officers' training schools, textbook and workbook publishing houses, test-preparation and distance-learning firms, online universities, Internet cheat busters, and the now ubiquitous computer classrooms of the schools.

During those years of the entrepreneurial race for the grading machine, 1988-2002, the official word from the composition field on automated scoring was barely audible. Hawisher et al.'s detailed *Computers and the Teaching of Writing in American Higher Education, 1979-1994* (1996) does not mention machine scoring. As late as 1993, William Wresch, as computer-knowledgeable as could be wished, summed up the "imminence of grading essays by computer" by saying there was no such prospect: "no high schools or colleges use computer essay grading . . . there is little interest in using computers in this way" (48). The first challenges to Wresch's pseudocleft "there is" came from people who had programs of their own to promote: Emil Roy and his Structured Decision System (Roy 1993), Ellis Page and his revamped Project Essay Grade (Page and Petersen 1995), and Hunter Breland and his WordMAP (Breland 1996). Not until Dennis Baron in 1998 and Anne Herrington and Charles Moran in 2001 did the ordinary run of college compositionists learn that grading essays by computer in fact was not imminent, it was here. Had they been so inclined they could have heard the Cassandra truth forty years earlier from Arthur Daigon who, in 1966, when only one program existed to rate student essays, got it precisely right: "In all probability, the first practical applications of essay grading by computer will be to tests of writing proficiency not returned to the writers, perhaps large scale testing of composition" (47).

Anyone who worked as a college writing teacher during the seventies, eighties, and nineties, as I did, will protest, saying that it is only right that our attention was directed at the use of computers for classroom instruction, not for housecleaning tasks such as placement. But it's too simple to say that composition was focused on instruction and not on evaluation, because we were focused on evaluation, too. Moreover, our traditional take on evaluation was very much in sympathy with automated scoring. The unpleasant truth is that the need the current machines fulfill is our need, and we had been trying to fulfill it in machinelike ways long before computers. So much so that when automated scoring actually arrived, it found us without an obvious defense. We've been hoist by our own machine.

The scoring machines promise three things for your money, all explicit in the home pages and the glossy brochures of industry automated-scoring packages: efficiency, objectivity, and freedom from drudgery. These three goals are precisely what writing teachers have been trying to achieve in their own practices by way of evaluation for a century. The goal of efficiency needs no brief. Our effort to reach the Shangri-la of fast response, quick return, and cheap cost can be seen in the discipline all the way from the periodic blue-ribbon studies of paper load and commenting time (average is about seven minutes a page) to the constant stream of articles proposing novel methods of response that will be quicker but still productive, such as my own "Minimal Marking" (Haswell 1983). Writing teachers feel work-efficiency in their muscles, but it also runs deep in our culture and has shaped not only industrialized systems of evaluation but our own ones as well (Williamson 1993, 2004). Objectivity also needs no brief, is also deeply cultural, and also shapes methods of writing evaluation from top to bottom. The student at the writing program administrator's door who wants a second reading brings an assumed right along with the essay and is not turned away. The few counterdisciplinary voices arguing that subjectivity in response to student writing is unavoidable and good (Dethier 1983; Markel 1991) are just that, few and counter to the disciplinary mainstream.

But drudgery is another matter. Surely writing teachers do not think of their work as drudgery. Do we think of ourselves as drudges?

Actually, we do. Long before computers we have used "drudgery" as a password allowing initiates to recognize each other. More literally, we often further a long tradition of college writing teachers separating off part of their work and labeling it as drudgery. In 1893, after only two years of teaching the new "Freshman English" course, professors at Stanford declared themselves "worn out with the drudgery of correcting Freshman themes" and abolished the course (Connors 1997, 186). My all-time favorite composition study title is nearly sixty years old: "A Practical Proposal to Take the Drudgery out of the Teaching of Freshman Composition and to Restore to the Teacher His Pristine Pleasure in Teaching" (Doris 1947). Forty-six years later, in *The Composition Teacher as Drudge: The Pitfalls and Perils of Linking across the Disciplines* (1993), Mary Anne Hutchinson finds new WAC systems turning writing teachers into nothing but copy editors, "Cinderellas who sit among the ashes while the content teachers go to the ball" (1). As these cites indicate (and scores in between), "drudgery" covers that menial part of our professional activity involved with marking papers. And it refers not to our true wishes but to lift-that-bale conditions imposed on us ("paper load"). When it comes to response, we are good-intentioned slaves. In 1983, with the first sentence to "Minimal Marking," I made the mistake of writing, in manuscript, that "many teachers still look toward the marking of a set of compositions with odium." When the piece appeared in print, I was surprised, though I should not have been, to find that the editor of *College English* had secretly changed "with odium" to "with distaste and discouragement." We really want to mark papers but want to do so with more efficiency, more objectivity, and less labor. As William Marling put it the next year, in explaining the motivation for his computerized paper-marking software while defending the continued need for teacher response, "The human presence is required. It is the repetitive drudgery I wanted to eliminate" (1984, 797; quoted by Huot 1996, which provides more evidence of the discipline's vision of computers as "a reliever of the drudgery of teaching writing," 236).

But long before computers, the drudgery we had been complaining about we had been trying to solve with machinelike or servantlike devices: labor-saving contraptions such as (in rough historical order) correction symbols, checklists, overhead projectors, rubber stamps, audiotapes; and cheap labor such as lay readers and student peer evaluators and teaching assistants ("the common experience for adjunct faculty remains drudgery," Soldofsky 1982, 865). So when the computer came along, we immediately saw it as the mechanical slave that could do our drudgery for us. Even as early as 1962, when cumbersome mainframe line editors were the only means of computer-aided response, decades before spell-checkers, word-processing AutoCorrect, and hypertext frames, Walter Reitman saw computers in this light: "Just as technology has helped to relieve the worker of much physical drudgery, so computer technology thus may free the teacher of much of his clerical drudgery, allowing him to utilize more of his energies and abilities in direct and creative contact with the individual student" (1962, 106). With a computer there would be no issue of odium, or even discouragement and distaste. The computer is an "unresentful drudge," as Henry W. Kucera put it five years later—Kucera, who had just programmed his machine to order 1,014,232 words by alphabet and frequency as it trudged through a digitized corpus of romance and western novels, government documents, religious tracts, and other mind-numbing genres (1967).

It was the discipline's special condition of drudgery that early visions of machine grading hoped, explicitly, to solve. Arthur Daigon, extolling Ellis Page's Project Essay Grade two years before the findings were published, said that it would serve "not as a teacher replacement but ultimately as an aid to teachers struggling with an overwhelming mass of paperwork" (1966, 47). Page himself wrote that it would "equalize the load of the English teacher with his colleagues in other subjects" (Page and Paulus 1968, 3). And three years later, Slotnick and Knapp imagined a computer-lab scenario where students would use a typewriter whose typeface could be handled with a "character reader" (scanner) so the computer could then grace their essays with automated commentary, thus relieving teachers "burdened with those ubiquitous sets of themes waiting to be graded" (1971, 75), unresentful commentary that, as Daigon hoped, would ignore "the halo effect from personal characteristics which are uncorrelated with the programmed measurements" (52). Later, in the 1980s, when the personal computer had materialized rather than the impersonal grader, interactive "auto-tutor" programs were praised because they never tired of student questions, spell-checkers and grammar-checkers were praised because they "relieved instructors of such onerous, time-consuming tasks as error-catching and proofreading" (Roy 1990, 85), autotext features of word-processing programs were praised because they could produce "boilerplate comments" for teachers "who face the sometimes soul-deadening prospect of processing yet another stack of student papers" (Morgan 1984, 6), and when research couldn't exactly prove that computers helped students write better essays at least the teacher could be sure that word-processing saved them from the "detested drudgery of copying and recopying multiple drafts" (Maik and Maik 1987, 11).

So when automated grading suddenly returned to the composition scene in the late 1990s, we should not have been entirely caught standing in innocence and awe. Didn't we get the drudge we were wishing for? For decades, on the one computing hand, we had been resisting automated rating in the name of mission and instruction, but on the other computing hand, we had been rationalizing it in the name of workload and evaluation. What right do we have to protest today when Nancy Drew's Web site argues that her Triplet Ticket software will turn "rote drudgery" into a "chance for quality learning" for both student and teacher (2004)?

BLACK BOXES

> *That [computers] are black boxes with mysterious workings inside needn't worry us more than it did the Athenian watchers of the planetarium of the Tower of Winds in the first century B.C. or the congregation that stood with Robert Boyle and wondered at the great clock at Strassburg. We need only be concerned with what goes on outside the box.*
>
> —Derek J. de Solla Price (at the 1965 Yale conference on Computers for the Humanities)

There is another machinelike method with which our profession has long handled the onus of evaluating student essays. That method is the system of formal assessment we use to admit and place students. There, often we have managed efficiency, objectivity, and drudgery in a very forthright way, by turning the task over to commercial testing firms such as the Educational Testing Service, ACT, and the College Board. In turn they have managed *their* issues of efficiency, objectivity, and drudgery largely by turning the task of rating essays over to the scoring apparatus called holistic rating. The holistic, of course, has long been holy writ among composition teachers, even when they didn't practice it themselves.

In this section I want to argue that with our decades-long trust in holistic scoring, we have again already bought into machine scoring.

The word *trust* (or should I say *ignorance?*) ushers in a complicating factor, in need of explication. Enter the black box.

In the parlance of cybernetics a "black box" is any construction, hardware or software, that one can operate knowing input and output but not knowing what happens in between. For most of us, the entire operation that takes place after we hit the "print" key and before we pick up the printout is a black box—we cannot explain what happens in between. But even expert computer scientists function—manage input and output—via many black boxes. For instance, they can handle computer glitches whose source they don't know with diagnostic tools whose operation they cannot explain. I want to argue the obvious point that for writing teachers commercial machine scoring is largely a black box and the less obvious point that for writing teachers, even for those who participate in it, even for those who help construct and administer it, holistic scoring is also largely a black box. Finally, I want to argue the conspiracy of the two. Even more so than machine scoring and teacher aids such as undergraduate peer graders and criteria check sheets, machine scoring and holistic scoring enjoy a relationship that is historically complementary, even mutually supportive, maybe even symbiotic. Investigating the black boxes of both will make this relationship clear.

What does it take to investigate a black box? I turn to Bruno Latour (1987), who applies the computer scientist's concept of the black box to the way all scientists practice their research. In so doing Latour offers some surprising and useful insights into black boxes in general. In the science laboratory and in science literature, a black box can be many things—a standard research procedure, a genetic strain or background used to study a particular phenomenon, a quality-control cutoff, the purity of a commercially available chemical, an unsupported but attractive theory. In essence, it is anything scientists take on faith. Latour's first insight is counterintuitive, that normal scientific advance does not result in gain but in loss of understanding of what happens between input and output, that is, in more rather than fewer black boxes. How can that be? Take the instance of a laboratory of scientists who genetically engineer a variant of the mustard plant *Arabidopsis thaliana* by modifying a certain gene sequence in its DNA. They know the procedure by which they modified the sequence. Later scientists obtain the seeds and use the resulting plants in their own studies, understanding that the gene structure is modified but quite likely unable to explain the exact procedure that altered it, though they will cite the original work in their own studies. Latour would point out that as the *Arabidopsis* variant is used by more and more secondhand experimenters, the obscurity of the original procedure will grow. Indeed, the more the original study is cited, the less chance that anyone will be inclined to open up that particular black box again. Familiarity breeds opacity.[3]

Latour's insight throws a startling light on scientific practices, which most people assume proceed from darkness to light, not the other way around. Ready support of Latour, though, lies right at hand for us: commercial machine scoring. The input is a student essay and the output is a rate stamped on the essay, and as the chapters in this volume demonstrate over and over, students, teachers, and administrators are accepting and using this output with the scantiest knowledge of how it got there. Proprietary rights, of course, close off much of that black box from outside scrutiny. A cat can look at a king, however, and we can mentally question or dispute the black boxes. What will happen? Latour predicts our request for enlightenment will be answered with more darkness: every time we try to "reopen" one black box, we will be presented with "a new and seemingly incontrovertible black box" (1987, 80). As we'll see, Latour's prediction proves right. But although our inquiry will end up with a Russian-doll riddle wrapped in a mystery inside an enigma, the direction in which one black box preconditions another is insightful. With current-day machine scoring, the black boxes always lead back to the holistic.

Start with an easy mystery, what counts as an "agreement" when a computer program matches its rate on an essay with the rate of a human scorer on the same essay. By custom, counted is either an "exact agreement," two scores that directly match, or an "adjacent agreement," two scores within one point of each other. But why should adjacent scores be counted as "agreement"? The answer is not hard to find. Whatever is counted as a "disagreement" or discrepancy will have to be read a third time. On Graduate Management Admission Test essays since 1999, using a 6-point scale, Educational Testing Service's e-rater has averaged exact matches about 52 percent of the time and adjacent agreements about 44 percent of the time (Chodorow and Burstein 2004). That adds up to an impressive "agreement" of 96 percent, with only 4 percent requiring a third reading. But only if adjacent hits are counted as agreement. If only exact agreement is counted there would have been 48 percent of the essays requiring a third reading. And that would lower interrater reliability below the acceptable rate.[4]

But the notion of reliability leaves us with a new black box (we'll set aside the issue of the cost of third readings). Why is high concordance among raters a goal rather than low concordance? Isn't multiplicity of perspectives good, as in other judgments on human performance with the complexity of essay writing? The answer is that the goal of the scoring is not trait analysis but a unitary rate. The machine is "trained" on the same traits that the human raters are, and both arrive at a single-number score, the machine through multiple regression and the humans through training in holistic scoring, where only five or six traits can be managed with efficiency. With e-rater, these traits include surface error, development and organization of ideas, and prompt-specific vocabulary (Attali & Burstein 2004).

More black boxes. We'll set aside the mystery of why the separate traits aren't scored, compared, adjusted, and reported separately (more cost?) and ask why these few particular traits were chosen out of the plentiful supply good writers utilize, such as wit, humor, surprise, originality, logical reasoning, and so on. Here there are a number of answers, all leading to new enigmas. Algorithms have not been developed for these traits—but why not? A trait such as "originality" is difficult to program—but any more difficult than "prompt-specific vocabulary," which requires "training" the program in a corpus of essays written on each prompt and judged by human raters? One answer, however, makes the most intuitive sense. The traits e-rater uses have a long history with essay assessment, and in particular with holistic scoring at Educational Testing Service. History is the trial that shows us these traits are especially important to writing teachers.

History may be a trial, but as Latour makes clear, it is also the quickest and most compulsive maker of black boxes. How much of that essay-evaluation trial was really just unthinking acceptance of tradition? Does anybody know who first determined that these traits are important, someone equivalent to our biological engineers who first created the genetic variant of *Arabidopsis*? Actually, it seems this black box can still be opened. We can trace the history of traits like "organization" and "mechanics" and show that at one time Paul B. Diederich understood what goes into them. It was 1958, to be precise, when he elicited grades and marginal comments from readers of student homework, statistically factored the comments, and derived these two traits along with four others, a factoring that was passed along, largely unchanged, through generations of holistic rubrics at the Educational Testing Services, where Diederich worked (Diederich 1974, 5–10). It's true that even in his original study, Diederich was trusting black boxes right and left. When one of the lawyers he used to read and comment on student writing wrote in the margin, "Confusing," Diederich could not enter into the lawyer's head to find out what exactly he meant before he categorized the comment as "organization" or "mechanics" (or even "language use" or "vocabulary") in order to enter another tally into his factoring formula. The human head is the final black box that, as good empirical engineers of the creature *Homo sapiens*, we can never enter, can know only through input and output. (For more about the influence of Diederich's study on later holistic rubrics, see Broad 2003; Haswell 2002)

Surely there is another enigma here that can be entered, however. Why does machine essay scoring have to feed off the history of human essay scoring? Why does ETS's e-rater (along with all the rest of the current programs) validate itself by drawing comparison with human raters? Why establish rater reliability with *human* scores? Why not correlate one program's rates with another program's, or one part of the software's analysis with another part's? If machine scoring is better than human scoring—more consistent, more objective—then why validate it with something worse? The answer is that, historically, the machine rater had to be designed to fit into an already existing scoring procedure using humans. Right from the start machine scoring was conceived, eventually, as a *replacement* for human raters, but it would have to be eased in and for a while work hand in hand with the human raters within Educational Testing Service's sprawling and profitable essay-rating operation. The Educational Testing Service, of course, was not the only company to splice machine scoring onto holistic scoring. Ellis Page reminds us that in 1965 his initial efforts to create computer essay scoring was funded by the College Board, and "The College Board," he writes, "was manually grading hundreds of thousands of essays each year and was looking for ways to make the process more efficient" (2003, 43). The machine had to learn the human system because the human system was already implemented. It is no accident that the criteria that essay-rater designers say their software covers are essentially Diederich's original holistic criteria (e.g., Elliott 2003, 72). Nor is it any accident that developers of machine graders talk about "training" the program with model essays—the language has been borrowed from human scoring procedures. (Is human rating now altering to agree with the machine corater? There's a black box worth investigating!)

Obviously at this point we have reached a nest of black boxes that would take a book to search and enlighten, a book that would need to study economic, cultural, and political motives as well as strictly psychometric ones. We've supported Latour's startling contention that "the more technical and specialized a literature is, the more 'social' it becomes" (1987, 62). Our inquiry has not led only into blind alleys, though, and we can now see one thing clearly about machine scoring. From the start it has been designed to emulate a method of human scoring, but not any old sort of method. It is of a very particular and I would say peculiar sort. That method is the holistic as practiced in commercial large-scale ventures, where a scorer has about two to three minutes and a four- to six-part rubric to put a single number between 0 and 4 or 0 and 6 on an essay usually composed unrehearsed and impromptu within less than forty minutes. Let's be honest about this. The case *for* machine scoring is not that machine decisions are equal or better than human decisions. The case *against* machine scoring is not that machine decisions are worse than human decisions. These are red-herring arguments. The fact is that so far machines have been developed to imitate a human judgment about writing that borders on the silly. The machine-human interrater reliability figures reported by the industry are something to be proud of only if you can be proud of computer software that can substitute one gimcrack trick for another. Ninety-six percent "agreement" is just one lame method of performance testing closely simulating another lame method. The situation is known by another cybernetic term, GIGO, where it little matters that we don't know what's in the black box because we do know the input, and the input (and therefore the output) is garbage.[5]

The crucial black box, the one that writing teachers should want most to open, is the meaning of the final holistic rate—cranked out by human or machine. In fact, in terms of placement into writing courses, we know pretty much the rate's meaning, because it has been studied over and over, by Educational Testing Service among others, and the answer is always the same, it means something not far from garbage. On the kind of short, impromptu essays levered out of students by ACT, Advanced Placement, and now the SAT exams, holistic scores have a predictive power that is pitiful. Regardless of the criterion target—pass rate for first-year composition, grades in first-year writing courses, retention from first to second year—holistic scores *at best* leaves unexplained about nine-tenths of the information needed to predict the outcome accurately.[6] No writing teacher wants students put into a basic writing course on this kind of dingbat, black-box prediction. But we walk into our classes and there they are, and this has been our predicament for decades, back when the score was produced by humans imitating machines and now when the score is produced by machines imitating humans.

So how complicit are we? For every writing teacher who counts surface features for a grade, assigns mastery-learning modules, or takes testing-firm scores on faith or in ignorance, there are many who respond to essays with the student's future improvement in mind, hold individual conferences, and spend hours reading and conferring over the department's own placement-exam portfolios. Across the discipline, however, there is an unacknowledged bent—one of our own particular black boxes—that especially allies us with the testing firms' method by which they validate grading software, if practice can be taken as a form of alliance. This bent consists of warranting one inferior method of writing evaluation by equating it with another inferior method. One accepts directed student self-placement decisions because they are at least as valid as the "inadequate data of a single writing sample" (Royer and Gilles 1998, 59), or informed self-placement because it replaces teachers who don't have enough time to sort records (Hackman and Johnson 1981), or inaccurate computer grammar-check programs because the marking of teachers is inconsistent, or boring auto-tutors because human tutors are subjective, or the invalidity of Page's machine scoring because of "the notorious unreliability of composition graders" (Daigon 1966, 47). One of the earliest instances of this bent is one of the most blatant (Dorough, Shapiro, and Morgan 1963?). In the fall of 1962 at the University of Houston, 149 basic-writing students received grammar and mechanics instruction in large "lecture" classes all semester, while 71 received the same instruction through a Dukane Redi-Tutor teaching machine (a frame-controlled film projector). At the end of the semester neither group of students performed better than the other on a correction test over grammar and mechanics: "the lecture and program instruction methods employed were equally effective" (8). Yet three pages later the authors conclude, "It is clear that . . . the programmed instruction was superior to the traditional lecture instruction." The tiebreaker, of course, is efficiency: "The programmed instruction sections handled more students more efficiently in terms of financial cost per student" (11). In the world of writing evaluation, two wrong ways of teaching writing can make a right way.[7]

DEI EX MACHINA

Sólo la difícil es estimulante

—José Lezama Lima

I began with an image of college writing teachers watching, helpless, as automated essay scoring invades higher education. I end with an agenda to release us from this deer-in-the-headlights stance.

First, we should not blame the commercial testing firms. They have filled a vacuum we abandoned, they have gravitated toward the profits, they have sunk their own R&D money into creation and testing of the programs, they have safeguarded their algorithms and prompts, they have marketed by the marketing rules, and they are reaping their well-earned payoffs—this is all in their entrepreneurial nature.

Second, that doesn't mean we should necessarily follow the path they have blazed. Nor does *that* mean that we should necessarily follow our own paths. With the assessment and evaluation of writing, probably the best rule is to be cautious about any route that has been tried in the past, and doubly cautious about programs that swear they have seen the Grail. Pick up again the forty-year history of writing evaluation at the University of Houston. I don't know how long they stuck with their 1961 "superior" Redi-Tutors, but in 1977 they saw student "illiteracy" as such a problem that they classified *all* their entering students as "remedial" writers and placed them into one of two categories, NP or BC. NP stood for "needs practice" and BC for "basket case." So they introduced an exit writing examination. In the first trial, 41 percent of African Americans and 40 percent of Hispanics failed. Despite these results and an ever-growing enrollment, they remained upbeat: "Writing can actually be taught in a lecture hall with 200 or more students. We are doing it" (Rice 1977, 190). In 1984 they installed a junior writing exam to catch "illiterate" AA transfers. They judged it a success: "The foreign students who used to blithely present their composition credits from the junior college across town are deeply troubled" (Dressman 1986-87, 15). But all this assessment consumed faculty and counseling time. So in 2003 they turned all their testing for first-year placement and rising-junior proficiency "exclusively" over to ACT's WritePlacer. They claim their problems are now solved. "WritePlacer Plus Online helps ensure that every University of Houston graduate enters the business world with solid writing skills," and "it also makes the university itself look even more professional"

(University of Houston 2003, 32). Other universities, I am suggesting, may want to postpone looking professional until they have looked professionally at Houston's model, its history, and its claims.

Third, not only do we need to challenge such claims, we need to avoid treating evaluation of writing in general as a black box, need to keep exploring every evaluative procedure until it becomes as much of a white box as we can make it. I say keep exploring because our discipline has a long history of Nancy Drew investigation into writing evaluation, longer than that of the testing firms. Our findings do not always concur with those of the College Board and Educational Testing Service, even when we are investigating the same box, such as holistic scoring. That is because our social motives are different, as Latour would be the first to point out. In fact, our findings often severely question commercial evaluation tactics. Stormzand and O'Shea (1924) found nonacademic adult writers (including newspaper editors and women letter writers) using the passive voice much more frequently than did college student writers, far above the rate red-flagged years later by commercial grammar-check programs; Freedman (1984) found teachers devaluing professional writing when they thought it was student authored; Barritt, Stock, and Clark (1986) found readers of placement essays forming mental pictures of the writer when decisions became difficult; my own analysis (Haswell 2002) snooped into the ways writing teachers categorized a piece of writing in terms of first-year writing-program objectives, and detected them ranking the traits in the same order with a nonnative writer and a native writer but assigning the traits less central value with the nonnative; Broad (2003) discovered not five or six criteria being used by teachers in evaluating first-year writing portfolios but forty-six textual criteria, twenty-two contextual criteria, and twenty-one other factors. This kind of investigation is not easy. It's detailed and time-consuming, a multiround wrestling match with large numbers of texts, criteria, and variables. Drudgery, if you wish a less agonistic metaphor. And dear Latour points out that as you challenge the black boxes further and further within, the investigation costs more and more money. To fully sound out the *Arabidopsis* variant may require building your own genetics lab. To bring e-rater construction completely to light may require suing the Educational Testing Service. "Arguing," says Latour, "is costly" (1987, 69). But without black-box investigations, we lack the grounds to resist machine scoring, or any kind of scoring. I second the strong call of Williamson (2004) for the discipline "to study automated assessment in order to explicate the potential value for teaching and learning, as well as the potential harm" (100).

Fourth, we need to insist that our institutions stop making students buy tests that do not generate the kind of outcomes right for our purposes. Here I am not saying anything new. For a quarter of a century now, researchers in composition have been showing that holistic scoring is not the best way to diagnose or record potential in student writing, yet potential is what placement in writing courses is all about. What I have been saying that may be new—at least the way it is disregarded suggests that it is new to quite a few people—is that from the start current machine scoring has been designed to be counterproductive for our needs. As I have said, the closer the programs get to traditional large-scale holistic rating—to this particular, peculiar method by which humans rate student essays—the less valid the programs are for placement.

Fifth, we need to find not only grounds and reasons but also concrete ways to resist misused machine scoring. Usually we can't just tell our administration (or state) to stop buying or requiring WritePlacer. Usually we can't just tell our administration we do not accept the scores that it has made our students purchase, even when we are willing to conduct a more valid procedure. For many of the powers that be, machine scoring is a deus ex machina rescuing all of us—students, teachers, and institution—from writing placement that has turned out to be a highly complicated entanglement without any clear denouement. The new scoring machines may have a charlatan look, with groaning beams and squeaking pulleys, but they work—that is, the input and the output don't create waves for management. So composition teachers and researchers need to fight fire with fire, or rather machine with machine. We need to enter the fray. First, we should demand that the new testing be tested. No administration can forbid that. Find some money, pay students just placed in basic writing via a commercial machine to retest via the same machine. My guess is that most of them will improve their placement. Or randomly pick a significant chunk of the students placed by machine into basic writing and mainstream them instead into regular composition, to see how they do. If nine-tenths of them pass (and they will), what does that say about the validity of the machine scores?

To these two modest proposals allow me to add an immodest one. We need to construct our own dei ex machina, our own golems, our own essay-analysis software programs. They would not be machine scorers but machine placers. They would come as close as machinely possible to predict from a pre-course-placement essay whether the student would benefit from our courses. Let's remember that the algorithms underlying a machine's essay-scoring protocol are not inevitable. Just as human readers, a machine reader can be "trained" in any number of different ways. Our machine placer would take as its target criterion not holistic rates of a student's placement essay but end-of-course teacher appraisals of the student's writing improvement during the actual courses into which the student had been placed. All the current methods of counting, tagging, and parsing—the proxes, as Page calls them—could be tried: rate of new words, fourth root of essay length, number of words devoted to trite phrases, percentage of content words that are found in model essays on the placement topics, as well as other, different proxes that are associated with situational writing growth rather than decontextualized writing quality. This machine placer would get better and better at identifying which traits of precourse writing lead to subsequent writing gain in courses. This is not science fiction. This can be done now. Then, in the tradition of true scholarship, let's give the programs free to any college that wants to install them on its servers and use them in place of commercial testing at $29 a head or $799 a site license. That will be easier even than hawking Kitchen Magicians. And then, in the tradition of good teaching, let's treat the scores not as single, final fiats from on high but embed them in local placement systems, systems that employ multiple predictor variables, retesting, course switching, early course exit, credit enhancement, informed self-placement, mainstreaming with ancillary tutoring—systems that recognize student variability, teacher capability, and machine fallibility.

Sixth, whatever our strategy, whatever the resistance we choose against the forces outside our profession to keep them from wresting another of our professional skills from out of our control, we have to make sure that in our resistance we are not thereby further debilitating those skills. We need to fight our own internal forces that work against good evaluation. Above all, we have to resist the notion of diagnostic response as rote drudgery, recognize it for what it is, a skill indeed—a difficult, complex, and rewarding skill requiring elastic intelligence and long experience. Good diagnosis of student writing should not be construed as easy, for the simple reason that it is never easy.

Here are few lines from a student placement essay that e-Write judged as promising (score of 6 out of possible 8) and that writing faculty members judged as not promising (they decided the student should have been placed in a course below regular composition). The prompt asks for an argument supporting the construction of either a new youth center or a larger public library.

> I tell you from my heart, I really would love to see our little library become a place of comfort and space for all those who love to read and relax, where we would have a plethora of information and rows upon rows of books and even a small media center. I have always loved our library and I have been one of those citizens always complaining about how we need more space, how we need more room to sit and read, how we need a bigger building for our fellow people of this community.
>
> But, I thought long and hard about both proposals, I really did, how nice would it be for young teens to meet at a local place in town, where they would be able to come and feel welcome, in a safe environment, where there would be alot less of a chance for a young adult of our community to get into serious trouble?

What is relevant here in terms of potential and curriculum? The careful distinctions ("comfort and space")? The sophisticated phrase "plethora of information"? The accumulation of topical points within series? The sequencing of rhetorical emphasis within series ("even")? The generous elaboration of the opposing position? The unstated antimony between "fellow people" and "young adult"? The fluid euphony of sound and syntactic rhythm? All I am saying is that in terms of curricular potential there is more here than the computer algorithms of sentence length and topic token-word maps, and also more than faculty alarm over spelling ("alot") and comma splices. Writing faculty, as well as machines, need the skill to diagnose such subtleties and complexities.

In all honesty, the art of getting inside the black box of the student essay is hard work. In the reading of student writing, everyone needs to be reengaged and stimulated with the difficult, which is the only path to the good, as that most hieratic of poets José Lezama Lima once said. If we do not embrace difficulty in this part of our job, easy evaluation will drive out good evaluation every time.