NOTES

CHAPTER 1 (MCALLISTER AND WHITE)

1. While the term "automated-essay scoring" (AES) is also frequently used, we prefer "computer-assisted writing assessment" because it more accurately reflects the current (and previous) state of this discipline. Virtually none of the work in computer-assisted writing assessment is automatic to the point of being autonomous yet, but rather requires numerous human-computer interactions; thus, computers are *assisting* in the partially automated writing-assessment process. It is also worth noting that "automation" does not necessarily involve computers. For example, Henry Ford and Elihu Root—Samuel Colt's lead engineer—both developed highly automated production systems long before the development of the computer.
2. A notable example of such articulate writing teachers are those who wrote the official position statement on computer-assisted writing assessment for the Conference on College Composition and Communication (2004):

> Because all writing is social, all writing should have human readers, regardless of the purpose of the writing. Assessment of writing that is scored by human readers can take time; machine-reading of placement writing gives a quick, almost instantaneous scoring and thus helps provide the kind of quick assessment that helps facilitate college orientation and registration procedures as well as exit assessments.
>
> The speed of machine-scoring is offset by a number of disadvantages. Writing-to-a-machine violates the essentially social nature of writing: we write to others for social purposes. If a student's first writing experience at an institution is writing to a machine, for instance, this sends a message: writing at this institution is not valued as human communication—and this in turn reduces the validity of the assessment. Further, since we can not know the criteria by which the computer scores the writing, we can not know whether particular kinds of bias may have been built into the scoring. And finally, if high schools see themselves as preparing students for college writing, and if college writing becomes to any degree machine-scored, high schools will begin to prepare their students to write for machines.
>
> We understand that machine-scoring programs are under consideration not just for the scoring of placement tests, but for responding to student writing in writing centers and as exit tests. We oppose the use of machine-scored writing in the assessment of writing. (798)

3. Ellis Page (2003) proposes a somewhat more broad set of categories into which critics of computer-assisted writing assessment fall: humanist (only humans can judge what humans have written); defensive (the testing environment is too complex for a computer to assess it correctly); and construct (computers can't accurately identify all the "important" variables that determine "good" writing) (51–52).
4. There are many examples of pre computer age stylistic analyses. See, for example, Charles Bally's *Traité de stylistique française* (1909), Caroline Spurgeon's *Shakespeare's Imagery and What it Tells Us* (1935), and Wolfgang Clemen's *Development of Shakespeare's Imagery* (1977). Wainer (2000) cites perhaps two of the most ancient

examples, one from around 2200 BCE, when a Chinese emperor implemented official testing procedures for his officials in a variety of disciplines including writing, and the second taken from the Hebrew Bible (Judges 12:4–6), in which people in a fleeing crowd were asked to say the word *shibboleth*; those who mispronounced it were suspected to be Ephraimites—a group prohibited from leaving—and were punished very harshly indeed (2).

5. Vantage Learning, the maker of IntelliMetric, includes this information on its Web site (2005a): "We take pride in our ability to develop and implement high-quality, large-scale online assessment programs. . . . Vantage Commercial's Language Recognizer™ uses natural language parsers to index documents in multiple languages, while our rule compilers parse the very specific rule specification languages used in our rule bases."
6. For refinements in Sager's work see *Foundational Issues in Natural Language Processing*, edited by Sells, Shieber, and Wasow (1991); *The Core Language Engine*, edited by Alshawi (1992); and *Machine Learning of Natural Language*, edited by Powers and Turk (1989).
7. Roy Davies (1989) recounts and expands upon Swanson's notion that "[k]nowledge can be created by drawing inferences from what is already known," for example, in published articles and books.
8. After Knowledge Analysis Technologies was purchased by Pearson Education, Landauer was named to his current position of executive vice president of Pearson Knowledge Technologies.

CHAPTER 2 (ERICSSON)

1. Although a foray into the meaning of *wisdom* is tempting here, I will resist the temptation and leave it to readers to ponder what Elliot might consider "wisdom" and how a computer program might attain or "internalize" that noble trait.
2. Their claim that "writing teachers are critical to the development of the technology because they inform us how automated essay evaluations can be most beneficial to students" (xv) is disingenuous in that it assumes that writing teachers accept this technology as something that could be beneficial to students—many teachers disagree with this assumption. This claim also leaves out writing scholars—the people who study writing and composition.
3. For proof of this claim, see McGee, chapter 5 in this volume.
4. Speculation on what happens to student writing when this "partner" is a computer is well worth consideration, but beyond the scope of this chapter.

CHAPTER 4 (HASWELL)

1. According to Dr. Nancy Drew's Web site, the Triplet Ticket proposes to make life easier for "today's over-burdened teachers." The promo repeats three classic warrants for machine scoring of student essays: eliminate human reader bias, reduce paper load, and provide immediate feedback. Next to a photograph filled with nothing but stacked essays is this text: "Assigning electronically graded essays as an instructional alternative counteracts the tendency for teachers to stop giving written essays because of grading overload" (Drew 2004). When I e-mailed her (August 2004), pointing out that her stated criteria for rating essays—spelling, sentence length, and essay length—could be calculated with count and find functions of any word-processing program, she answered that the statement was a mistake of her Web page writer, that there were other criteria, and that she could not divulge them because of a pending patent.
2. In 1985, Quintilian Analysis required the student or the teacher to enter the essay via line editing (no word wrap) and to insert special coding characters marking end of paragraph and parts of speech. The output included gentle advice worthy of Mr. Chips: "Your sentences run to the short side, typical of popular journalism

or writing for audiences unwilling to cope with longer sentence constructions. Are you using such short sentences for some particular effect? Are you trying to outdo Hemingway?" It sold for $995. The author, Winston Weathers, is better known for his theory and pedagogy of alternative styles.

3. At this point in reading my essay my daughter Elizabeth, a plant biologist, had had enough, commenting: "Familiarity also breeds efficiency. No scientist wants or needs to test every claim that is published by others; trust in the work of others is required for scientific advances." Agreed—and a truth that applies to all labor, not just scientific, including the labor of writing teachers. So the issue is not just what's the efficiency, but whom do you trust and when do you question. The example of *Arabidopsis* is my own, by the way, not Latour's, whose analysis of many other black boxes is hard to beat (21–62).
4. Other programs achieve similar rates. IntelliMetric's performance is exact agreement 57 percent of the time, adjacent agreement 41 percent (Elliott 2003). That's a very profitable 2 percent third-reading rate with the usual definition of "agreement," and a costly 43 percent third-reading rate with an exact agreement definition. In selling the software today, while the standard magic formula is "the machine agrees with human raters as well as human raters agree with each other," some promoters go further. IntelliMetric, according to Scott Elliot, "will typically outperform human scorers" (75), and Ellis Page makes the same claim for Project Essay Grade (Page and Petersen 1995). They can say that because their machine scores correlate better with the *mean* score of a group of raters than any one of the rater's scores do with that average. They don't say what is so good about an average score. Another black box.
5. GIGO: garbage in, garbage out. Again, in 1966 Arthur Daigon got it right, or almost right. After his prediction that computer grading would first be used in "large scale testing of composition," he shrewdly added that this "would merely require simulation of the single evaluative end product of enlightened human judgment. Is the composition *unacceptable*, *fair*, *good*, or *excellent*?" (47). The question is whether reducing a piece of writing to a "single evaluative end product" (i.e., rate), with a discrimination no more informative than 1, 2, 3, and 4, constitutes human judgment that one can call "enlightened."
6. "Pitiful" is not an exaggeration. Technically speaking, holistic score explains around 9 percent of the total variance of the target criterion. That's an average of many studies (for a review, see McKendy 1992). Educational Testing Service's own researchers have improved this predictive power by creating optimal conditions, and then only minimally. The best Breland et al. (1987) could achieve was 33 percent on essays written at home on announced topics. In the customary short, impromptu, sit-down conditions of Educational Testing Service testing, Brent Bridgeman (1991), another Educational Testing Service researcher, found that a holistically scored essay added zero to a prediction of freshman grades, a prediction formula combining high school GPA, SAT scores, and a multiple-choice test of writing-skill knowledge. For Educational Testing Service this is truly being hoist by your own petard. The higher Educational Testing Service achieves a correlation with machine scores and human holistic scores, the less grounds—by their own research—they have to argue that machine scores should serve for placement. And what's true of Educational Testing Service is equally true of the other automatic rating enterprises. It's no surprise that Shermis and Burstein's *Automated Essay Scoring* (2003), that book-length argument for machine scoring from the industry side, reports not one completed study of the *instructional* validity of machine scores. On the crucial distinction between old-fashioned test validity (to which commercial validation of machine scoring sticks) and current contextual or instructional or decision validity, see Williamson 2004.
7. The art of validating one poor method of writing assessment by equating it with another poor method has been long practiced on the commercial side. For a typical example, see Weiss and Jackson's conclusion to their College Board study (1983)

that found an indirect measurement of writing proficiency, the Descriptive Tests of Language Skills, predicting college writing-course performance (final grade and post-essay) as badly as did a pre-essay. The predictive coefficient for all was terrible, around .4, but they still say, "In fact, each of the Descriptive Tests of Language Skills scores was found to predict posttest essay scores about as well as pretest essay scores did and somewhat better than self-reported high school English grades did. Thus, these results lend support to the use of the Descriptive Tests of Language Skills as an aid in making decisions about the placement of students in introductory level college composition courses" (8). On the instructional side, the rationale that validates a new computer-aided method of instruction because it is no worse than a previous computerless method is standard in defense of online distance-learning courses. See Russell 1999.

CHAPTER 5 (MCGEE)

1. This and subsequent quotations were taken from the Knowledge Analysis Technologies Web site in April 2001. The site has since undergone a major revision, and while some of the promotional copy from the earlier version persists unchanged, the seemingly hyperbolic claims about understanding "the meaning of written essays" no longer appear.
2. In a 2002 memo to the provost requesting funds to administer the test, an economics professor asserted that the "ETS test is a cheap and effective way to get reliable third-party assessments of our students' writing skills" (Vandegrift).
3. In his "Apologia for the Timed Impromptu," Edward White (1995) concedes many weaknesses of timed impromptus and lists the kinds of advanced composing skills that short impromptu essay tests are "unlikely . . . [to] provide us with much useful information about" (34). Consequently, invoking sophisticated text-analytical approaches to the scoring of such essays seems like analytical overkill.
4. That belief was later ratified when I had an opportunity to compare the scores Criterion awarded to the essays by students whose "diagnostic essays" I had already scored holistically. There was one malfunction, one higher than expected, one lower than expected, and the remaining thirty-three matched very closely with the scores I had awarded.
5. In a 1950 paper, Alan Turing proposed a test of a computer's ability to produce humanlike conversation, asserting that a machine that could pass such a test deserved to be called intelligent. One source describes the test as follows: "a human judge engages in a natural language conversation with two other parties, one a human and the other a machine; if the judge cannot reliably tell which is which, then the machine is said to pass the test" ("The Turing Test" 2005)
6. Cynics may find something distasteful in the testing corporations' assertion that a fair assessment of their machines depends upon students having made a "good faith effort " and suggest that students attempting to psyche out a scoring machine did not initiate the cycle of bad faith.
7. My understanding of cohesion is informed largely by Joseph Williams's *Style: Ten Lessons in Clarity and Grace* (2000). Having used that text on multiple occasions, I found his treatment of the related principles of cohesion and coherence particularly useful when teaching college students who crafted decent sentences but didn't do enough to ease the cognitive burden on readers attempting to make meaning of new information.
8. The original essay included multiple references to Huey Long, whose name I considered reversing to "Huey Short," but decided, instead, to revise it to Huey Newton.

CHAPTER 6 (JONES)

1. The relationships between ACCUPLACER, WritePlacer *Plus*, Vantage Learning, and IntelliMetric can be confusing. ACCUPLACER is the company that

purveys online placement testing. WritePlacer *Plus* is the essay test portion within ACCUPLACER. It uses the technology called IntelliMetric, created by the Vantage Learning Company.

2. All student names are pseudonyms.
3. Fifty-six essays from 2004 were chosen to represent students who had the widest possible range of sentence skills and reading scores. Eight-two essays from 2002 were random in that I just picked essays from the first ten students in an alphabetized list of students within each score level. These two batches comprise the "more or less randomly picked essays." Another eleven essays had already been identified as potentially anomalous by Nancy Enright and me.
4. All analyses were significant at the $p < .001$ level, meaning that the odds of this finding having resulted from a random distribution are less than one in a thousand.
5. It may be that the capacity of IntelliMetric to distinguish more subtle aspects of writing has improved some since 2002. The variance in essay scores explained by length alone was 90 percent in 2002. In 2003, the percentage was 82 percent, while in 2004 the percentage was 84 percent.

CHAPTER 7 (HERRINGTON AND MORAN)

1. In addition to products for writing placement at college entry level, the products include programs for high-stakes statewide assessment: for example, Vantage Technologies lists the Oregon Department of Education as a client for its Technology Enhanced Student Assessment, a "high-stakes statewide assessment system," also the Pennsylvania Department of Education for its statewide assessment system, and a product for CTB/McGraw Hill for "direct assessment for use by K–12 institutions" (Vantage Learning 2005a). The capability of standardized assessment and record keeping to track that assessment become bases for marketing products for use in the classroom for assessing writing, tracking performance, and even providing feedback on writing: for example, Vantage Learning's Learning Access, a comprehensive program marketed as helping K–9 teachers "meet the challenges of No Child Left Behind," MY Access! an "Online Writing Development Tool," developed in partnership with the Massachusetts Department of Education and designed for classroom use, and ETS's Criterion, for "online writing evaluation in college classrooms." Another type of product, in which KAT has taken the lead, is aimed at content assessment, using the Intelligent Essay Assessor program. KAT's Web site lists such clients as the U.S. Air Force Research Laboratories for a Career Map Occupational Analysis Program; Prentice Hall for a companion Web site to the text *Keys to Success*; the University of Colorado for the Colorado Literacy Tutor, designed for "individualized, computer-aided reading instruction"; and Florida Gulf Coast University for an automated essay-assessment program for a large online course, Understanding the Visual and Performing Arts.

CHAPTER 14 (ROTHERMEL)

1. This promotional brochure has been replaced by the online Product Sheet (Vantage Learning 2004d).