Should a Computer Grade Your Essays?
CASE STUDY

Would you like your college essays graded by a computer? Well, you just might find that happening in your next course. In April 2013, EdX, a Harvard/MIT joint venture to develop massively open online courses (MOOCs), launched an essay scoring program. Using artificial intelligence technology, essays and short answers are immediately scored and feedback tendered, allowing students to revise, resubmit, and improve their grade as many times as necessary. The non-profit organization is offering the software free to any institution that wants to use it. From a pedagogical standpoint—if the guidance is sound—immediate feedback and the ability to directly act on it is an optimal learning environment. But while proponents trumpet automated essay grading's superiority to students waiting days or weeks for returned papers—which they may or may not have the opportunity to revise—as well as the time-saving benefit for instructors, critics doubt that humans can be replaced.

In 2012, Les Perelman, the former director of writing at MIT, countered a paper touting the proficiency of automated essay scoring (AES) software. University of Akron College of Education dean, Mark Shermis, and co-author, data scientist Ben Hamner used AES programs from nine companies, including Pearson and McGraw-Hill, to rescore over 16,000 middle and high school essays from six different state standardized tests. Their Hewlett Foundation sponsored study found that machine scoring closely tracked human grading, and in some cases, produced a more accurate grade. Perelman, however, found that no direct statistical comparison between the human graders and the programs was performed. While Shermis concedes that regression analysis was not performed—because the software companies imposed this condition in order to allow him and Hamner to test their products—he unsurprisingly accuses Perelman of evaluating their work without performing research of his own.

Perelman has in fact conducted studies on the Electronic Essay Rater (e-rater) developed by the Educational Testing Service (ETS)—the only organization that would allow him access. The e-rater uses syntactic variety, discourse structure (like PEG) and content analysis (like IEA)and is based on natural analysis to linguistic features like argument formation and syntactic variety to determine scores, but also gives weight to vocabulary and topical content. In the month granted him, Perelman analyzed the algorithms and toyed with the e-Rater, confirming his prior critiques. The major problem with AES programs (so far) is that they cannot distinguish fact from fiction. For example, in response to an essay prompt about the causes for the steep rise in the cost of higher education, Perelman wrote that the main driver was greedy teaching assistants whose salaries were six times that of college presidents with exorbitant benefits packages including South Seas vacations, private jets, and movie contracts. He supplemented the argument with a line from Allen Ginsberg's "Howl," and received the top score of 6. The metrics that merited this score included overall length, paragraph length, number of words per sentence, word length, and the use of conjunctive adverbs such as "however" and "moreover." Since computer programs cannot divine meaning, essay length is a proxy for writing fluency, conjunctive adverb use for complex thinking, and big words for vocabulary aptitude.

Program vendors such as Pearson and Vantage Learning defend these parameters, asserting that they are highly correlated. Good writers have acquired skills that enable them to write more under time constraints; they use more complex vocabulary, and they understand how to introduce, interrupt, connect, and conclude complex ideas—the jobs of conjunctive adverbs. AES programs also recognize sentence fragments and dock students for sentences that begin with "and" or "or." However, professional writers know how to employ both to great effect. Perelman and a newly formed group of educators, Professionals Against Machine Scoring of Student Essays in High-Stakes Assessment, warn that writing instruction will be dumbed down to meet the limited and rigid metrics machines are capable of measuring.

The productivity gains from using automated essay-grading software will undoubtedly take away some of the jobs of the graders hired by the standardized test companies. Pearson, for example, ostensibly pays its graders between $40 and $60 per hour. In

and 30 essays—that's two to three minutes (and dollars) per essay. Clearly graders must use some type of shorthand metrics in order to score this quickly, but at least they can recognize as false the statement that on July 4, 2013, the United States observed its 2,013th birthday, even if it is contained in a well-constructed sentence. While the e-Rater can score 16,000 essays in 20 seconds, it cannot make this distinction. And presumably, a 716-word essay containing multiple nonsense sentences will not receive a 6 from a human grader while a 150-word shorter, factual, well-reasoned essay scores a 5, as Perelman was able to demonstrate.

ETS, developer of the SAT, GRE, Praxis, and K-12 standardized tests for multiple states, counters that the e-Rater is not replacing human graders in high-stakes tests; it is supplementing them. Essays are scored by both human and machine and when the scores do not match, a second human breaks the impasse. Furthermore, they posit that the test prep course Perelman developed to teach students how to beat AES software requires higher-order thinking skills—precisely those the tests seek to measure. Thus, if students can master Perelman's techniques, they have likely earned their 6. Pearson adds that its Intelligent Essay Assessor is primarily a classroom tool, allowing students to revise their essays multiple times before turning them in to a teacher to be graded. But for many states looking to introduce writing sections to their battery of K-12 standardized tests, and for those that abandoned the effort due to the cost, eliminating graders altogether will make them affordable. And the stakes are not insubstantial for failure to achieve passing grades on state standardized tests, ranging from retesting, to remedial programs, to summer school, to non-promotion.

The free EdXtool appears to be more sophisticated than some vendor offerings in that it is "trainable" with at least some ability to develop grading standards and to adapt to grading preferences. First, instructors grade 100 essays or essay questions, and these are input to the program. Using these guidelines, the tool develops customized grading metrics and follows the scoring method preferred by the instructor, either a numerical system or letter grade. As noted by Shermis, in many lesser-ranked colleges than those of the critics, classes are now so large as to render comprehensive writing feedback infeasible. Moreover, at top universities, the instructional level is higher with fewer students in need of remediation. Down in the educational trenches, a tool that can adequately simulate human scoring, with no greater variation than that seen from instructor to instructor, and that provides immediate guidance, is a welcome addition to the instructional toolbox. But as demands on instructor's time decrease, will university administrators push staff cutbacks to meet budgetary constraints? Will fewer and fewer instructors be teaching more and more students?

As MOOC and AES proliferate, the answer is: most likely. EdX is quickly becoming controversial in academic circles. Presently, its course offerings are free and students earn a certificate of completion, but not course credit. To become self-sustaining, however, the non-profit plans to offer its MOOC platform as a "self-service" system, which faculty members can use to develop courses specifically branded for their universities. EdX will then receive the first $50,000 in revenue generated from the course or $10,000 for a recurring course. Thereafter, revenue will be split 50–50 between the university and EdX. A second revenue-generating model offers universities "production help" with course development, charging them $250,000 for a new course and $50,000 each term the course is offered again. If a course is successful, the university receives 70% of the revenue, as long as EdX has been fully compensated for any self-service courses. However, in order to generate enough revenue to share with its 12 university partners, which now include University of California, Berkeley, Wellesley, Georgetown, and the University of Texas, a licensing model is likely. Tested at no charge at San Jose State University in 2012, an EdX MOOC served as the basis for a blended online engineering course. The enriched curriculum resulted in an increased passing rate from 60% to 91%. If course licensing becomes the key revenue stream, Anant Agarwal, the electrical engineer president of EdX, foresees this happening in closed classrooms with limited enrollment.

But some members of the San Jose State faculty are nonetheless alarmed. When a second EdX MOOC, JusticeX, was considered, the Philosophy department sent a sharply-worded letter addressed to Harvard course developer, Michael Sandel, but actually leveled at university administrators. Asserting that the department did not have an academic problem in need of remediation and was not lacking faculty to teach its equivalent course, it did not shy from attacking the economic motives behind public universities' embrace of MOOCs. The authors further asserted that MOOCs represented a decline in educational quality and noted the irony involved when a social justice course was the vehicle for perpetrating a social injustice—a long-term effort to "dismantle departments and replace professors."

Sandel's conciliatory response expressed his desire to share free educational resources, his aversion to undercutting colleagues, and a call for a serious debate at both EdX and in the higher education community.

Other universities are similarly pushing back, against both EdX and other new MOOC ventures such as Coursera and Udacity, founded by Stanford faculty members. MOOCs and AES are inextricably linked. Massive online courses require automated assessment systems. And both Coursera and Udacity have expressed their commitment to using them due to the value of immediate feedback. Amherst College faculty voted against joining the EdX consortium. Duke University faculty members thwarted administration attempts to join nine other universities and educational technology company 2U in a venture to develop a collection of for-credit undergraduate courses.

But EdX was founded by two of the most prominent universities in the United States, has gathered prestigious partners, and is already shaping educational standards. Stanford, for one, has decided to get on board; it adopted the OpenEdX open-source platform and began offering a summer reading program for freshman and two public courses in the summer of 2013. Stanford will collaborate with EdX on the future development of OpenEdX and will offer both public and university classes on it.

So while Professor Perelman jokes that his former computer science major students could develop an Android app capable of spitting out formulaic essays that would get a 6 from e-Rater, cutting humans completely out of the equation, he knows that serious issues are in play. What educational outcomes will result from diminishing human interaction and input? Will AI develop to the point that truth, accuracy, effective organization, persuasiveness, argumentation and supporting evidence can be evaluated? And how many more jobs in education will disappear as a result?

Sources: Caralee J. Adams, "Essay-Grading Software Seen as Time-Saving Tool," Education Week, March 10, 2014; www.writetolearn.net, accessed August 1, 2014; www.humanreaders.org, accessed July 28, 2014; Michael Gonchar, "How Would You Feel About a Computer Grading Your Essays?" *New York Times*, April 5, 2013; John Markoff, "Essay-Grading Software Offers Professors a Break," *New York Times*, April 4, 2013; Ry Rivard, "Humans Fight Over Robo-Readers," Inside Higher Ed, March 15, 2013; David Rotman, "How Technology Is Destroying Jobs, MIT Technology Review, June 12, 2013; Randall Stross, "The Algorithm Didn't Like My Essay, *New York Times*, June 9, 2012; Michael Winerip, "Facing a Robo-Grader? Just Keep Obfuscating Mellifluously," *New York Times*, April 22, 2012; Paul Wiseman, Bernard Condon, and Jonathan Fahey, "Can smart machines take your job? Middle class jobs increasingly being replaced by technology," The Associated Press, January 24, 2013;and "San Jose State University Faculty Pushes Back Against EdX," Inside Higher Ed, May 3, 2013.

CASE STUDY QUESTIONS

2-13 Identify the kinds of systems described in this case.

2-14 What are the benefits of automated essay grading? What are the drawbacks?

2-15 Can automated essay grading replace a human grader? Why or why not?

2-16 What management, organization, and technology factor should be considered when deciding whether to use AES?

2-17 Would you be suspicious of a low grade you received on a paper graded by AES software? Why or why not? Would you request a review by a human grader?

MyMISLab

Go to **mymislab.com** for Auto-graded writing questions as well as the following Assisted-graded writing questions.

2-18 Identify and describe the capabilities of enterprise social networking software. Describe how a firm could use each of these capabilities.

2-19 Describe the systems used by various management groups within the firm in terms of the information they use, their outputs, and groups served.