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 language processing technology. It applies statistical 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 that hour, a grader is expected to score between 20