The Washington Post

Answer Sheet

IS MIT researcher being censored by Educational Testing Service?

By Valerie Strauss October 24, 2014

Les Perelman is a research affiliate in the Comparative Media Studies/Writing program at the Massachusetts Institute of Technology, where he spent years as a director of undergraduate writing. He researches ways to improve student writing and has been a vocal critic of the automated grading of essays. Perelman, along with MIT and Harvard students, designed the Basic Automatic B.S. Essay Language Generator, or Babel, a machine built to prove that essay-grading software is very limited in its ability to find meaning or check the accuracy of a piece of writing. In this post, Perelman says that the Educational Testing Service, the world's largest private nonprofit educational testing and assessment organization, is censoring him in his effort to test a product ETS is selling to schools. ETS denies it in a response following Perelman's piece.

By Les Perelman

The Educational Test Service (ETS) won't let me continue to test a product that they are trying to sell to schools and colleges across America. Specifically, the company will not allow me access to the Automated Scoring Engine (AES) unless I agree to let them censor my findings.

All I want to do is test the claim by ETS that the feedback their automated essay scoring engine gives students is more precise than that available to anyone through Microsoft Word. Their website says: "The Microsoft® Word Spelling and Grammar tool can provide writers with a quick analysis of common errors. However, the Criterion service, as an instructional tool used to improve writing, targets more precise feedback."

I submitted a proposal to test this claim to ETS. Previously, access to Criterion was easy to obtain. In 2012, in a series of experiments publicized in The New York Times , I demonstrated that Criterion was oblivious to factual errors and intentional incoherence. In fact, the scoring system preferred long pretentious language and verbosity.

I also discovered that Criterion's feedback was often inaccurate and sometimes just plain wrong. It categorizes perfectly appropriate uses of the definite article as a "missing or extra article," told me that the phrase "opinions about a film" contained a preposition error, and was almost always incorrect in identifying the thesis sentence of an essay. In addition, it chided me for writing a paragraph that had only three sentences. Nevertheless, I was quoted in The New York Times praising ETS for allowing me access. In the same article, David Williamson, the senior research director for the Assessment Innovations Center, was quoted as stating, "At E.T.S., we pride ourselves in being transparent about our research."

Well ETS is no longer so transparent. Earlier this year, I asked for access to *e-rater*, Criterion's scoring engine as part of a series of experiments to show that computer generated nonsense could receive high scores from Automated Essay Scoring (AES) computers. I even imagined that the nonsense generator could become a mobile app. ETS turned down my request, stating that I was developing a commercial product. In response, I pledged not to commercialize the research and appealed their decision. A one-sentence email informed me that the appeal was denied.

Three very smart undergraduates, two at MIT and one at Harvard, developed the computer gibberish tool, which we dubbed the Basic Automated BS Essay Language Generator or BABEL Generator. It works spectacularly well in producing nonsense that received high scores from various AES machines.

BABEL even works well with ETS' *e-rater*. Though ETS would not give me direct access to its scoring engine, ETS allows prospective test takers to take practice Graduate Record Exam (GRE) essays that are instantly graded by e-rater for $13. I have bought a number of these packages and have used the BABEL Generator to produce essays that consistently receive scores of 5 and 6 on a 1-6 point scale on each of the two writing assignments that comprise the writing portion of the GRE.

For example, one essay containing this language:

> Competition which mesmerizes the reprover, especially of administrations, may be multitude. As a result of abandoning the utterance to the people involved, a plethora of cooperation can be more tensely enjoined. Additionally, a humane competition changes assemblage by cooperation. In my semiotics class, all of the agriculturalists for our personal interloper with the probe we decry contend.. . .

is followed by these canned comments that the essay:

- articulates a clear and insightful position on the issue in accordance with the assigned task
- develops the position fully with compelling reasons and/or persuasive examples
- sustains a well-focused, well-organized analysis, connecting ideas logically
- conveys ideas fluently and precisely, using effective vocabulary and sentence variety
- demonstrates superior facility with the conventions of standard written English (i.e., grammar, usage, and mechanics) but may have minor errors

My success with the BABEL Generator spurred me to test the efficacy of the bold claims made for classroom applications of Automated Essay Scoring. Not only were computers going to test students but they were going to teach them as well, allowing students to write more and in classes in which an over-burdened instructor had 35 students and could just look at final drafts. As a Writing Program Administrator for over 30 years, I laugh at this naiveté, knowing that if such a technology were deployed, the most likely result would be that a superintendent or dean would double the class size to 70 students. In addition, a recently published article by two researchers showed that the Criterion was substantially less accurate than expert human scorers in identifying errors in papers written by advanced non-native English speakers. [Semire Dikli, Susan Bleyle, Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback?, *Assessing Writing*, Vol.22, October 2014, Pages 1-17.] Indeed, it also confirmed the patterns I found in my earlier studies that over forty percent of the errors identified by Criterion, especially those regarding articles, were not errors at all. While this defect may just be annoying to native speakers, it is devastating to those learning English.

I submitted a detailed proposal to compare the accuracy of Criterion to that of the Microsoft® Word Spelling and Grammar tool. I would conduct the study with a colleague from MIT who has a Ph.D. in linguistics from MIT and who worked with Noam Chomsky and Morris Halle, the founders of modern linguistics.

Instead of the easy access I received two years ago, I received a long email from Chaitanya Ramineni, a research scientist in the Innovations in the Development and Evaluation of Automated Scoring (IDEAS) Group that included some disturbing provisions. Unlike my previous experiences with Criterion, I would not be allowed access to Criterion but would be required to give the data to ETS for them to process and return to me. Even more disconcerting, was provision 3C:

> All presentations/manuscripts must be submitted for ETS review at least two weeks prior to public dissemination. ETS will retain the right to comment on the article to correct any errors, and as a result of the review, ETS can require that the ETS name, the Criterion name, and any identifying information about the particular company/product be removed from the publication / presentation / public dissemination (article, blog, etc.).

My reply to ETS labeled this censorship.

Dr. Ramineni responded, “This provision is a common policy/practice at ETS for external researchers who are not working directly with ETS staff.” This stance was reiterated in a phone conference I subsequently had with Dr. Ramineni, David Williamson, and several silent ETS executives and then repeated by the same group in another phone conference in which I was absent to the Chair of my professional organization a few weeks later.

Over the next few months, I discovered that the provisions ETS had told me were common practice were not consistently applied. Around the same time, another researcher had applied to use Criterion and had no problem gaining access. Indeed, ETS even provided her with training on how to use the latest release. The Criterion Non-Commercial Research Software License agreement [attached] she was asked to sign contained no language censoring content. And, during a phone call with Dr. Ramineni, she was asked out-of-the-blue if she was working with Les Perelman. She is not.

All I want to do is what organizations like Consumers Union and the Underwriters Laboratory do all the time: determine 1) if an advertised product meets its claims and 2) whether or not it is defective. Considering that the product in question is being used by school children and bought largely through public funds, free access should be limited solely to concerns about intellectual property. Yet ETS will not allow me access.

ETS is not alone. Pearson Educational Technologies wouldn’t even reply to my request to test their WriteToLearn® software, and Peter Foltz, a Pearson Vice President, was quoted in the 2012 New York Times article as justifying Pearson’s refusal to give me access to their product because “He wants to show why it doesn’t work.”

Although no company should prevent consumers from discovering whether or not its products work, ETS’s refusal is particularly alarming. ETS is a tax-exempt 501(c)(3) non-profit corporation. ETS claims this status as a non-profit educational institution, that, according to its charter, among other activities, encourages research in major areas of assessment.

The IRS, however, puts additional requirements on educational institutions:

The method used by an organization to develop and present its views is a factor in determining if an organization qualifies as educational within the meaning of section 501(c)(3). The following factors may indicate that the method is not educational.

1. The presentation of viewpoints unsupported by facts is a significant part of the organization's communications.
2. The facts that purport to support the viewpoint are distorted.

I am trying to verify the factual accuracy of important educational claims made by ETS but the company is trying to prevent me from doing so. I hope that someone at the IRS is reading this.

Here's a response from ETS Corporate Spokesperson Thomas Ewing:

> Our policy is not censorship, but is actually designed to be as transparent as possible while still protecting ETS from dissemination of erroneous information.
>
> What we have is a difference of viewpoint on policies relating to how reports which draw upon ETS-provided data can be published. We encouraged Mr. Perelman to use Criterion® for research purposes, but with the proviso that ETS be given the opportunity to review the results before they are published. The purpose of this policy is to ensure that there are no factual inaccuracies in public materials.
>
> For example, if an external researcher were to find a correlation between Criterion scores and performance in college they may

claim that Criterion is good for college admissions testing. However, this would be a misuse of Criterion and so ETS would be obligated to discourage this kind of claim.

Such a right of review for public material is a standard requirement for all such requests and something which other researchers accept, including our own, who submit papers for internal peer review before publication. We require that all presentations or manuscripts that rely upon ETS data be submitted for ETS to review at least two weeks prior to public dissemination. Obviously we have the right to comment on the research and to correct any errors or misrepresentations as a result of that review.

If we find the results of any study to be flawed or incorrect, the author can either correct the problems or they can remove any reference to ETS and replace them with general descriptions of the data that do not identify ETS as the source. This still allows the author to publish the results. This policy is progressive and more permissive than many other organizations who simply don’t allow external researchers to study their data, or publish on the basis. Mr. Perelman refused to accept this requirement and that’s where it stands.

Valerie Strauss covers education and runs The Answer Sheet blog. Follow @valeriestrauss

The Post Recommends

Trump gives his hard-line campaign promises a more moderate tone in address to Congress

Trump had an air of seriousness and revealed flashes of compassion as he broadly outlined a sweeping agenda to rebuild the country.

Big surge for military in Trump budget, big cuts elsewhere

President Donald Trump is proposing a huge $54 billion surge in U.S. military spending for new aircraft, ships and fighters in his first federal budget while slashing big chunks from domestic programs and foreign aid to make the government “do more with less.”

Big surge for military in Trump budget, big cuts elsewhere

President Donald Trump is proposing a huge $54 billion surge in U.S. military spending for new aircraft, ships and fighters in his first federal budget while slashing big chunks from domestic programs and foreign aid to make the government “do more with less.”