MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Comparative Media Studies / Writing

77 Massachusetts Avenue
Cambridge, MA 02139

Room 14E-304
EMail: perelman@mit.edu

Leslie C. Perelman, Ph.D.
Research Affiliate
Director of Writing Across the Curriculum, Retired

May 24, 2016

Response to the letter from the Dr. Jack R. Smith, the Interim State Superintendent of the Maryland State Department of Education.

.

I will address specific claims in Dr. Smith's letter as well as making some additional observations:

1. **"Maryland has been using automated scoring on the extended constructed response questions included in the science assessments for the past decade."** Machine scoring of short answer (one paragraph or less) questions for content appears to be fairly accurate and reliable. There is a major difference, however, between a machine looking for two or three key terms and parsing the relationship among them, and a machine attempting to score grammar much less coherence, development, argumentation, and other traits associated with English Language Arts (ELA).

2. **"Studies have shown that automated scoring is as accurate as human scoring and more consistent and efficient."** Because writing is communication from one mind to another mind, unless a machine can pass the Turing Test's "Imitation Game," a claim that machines are as accurate as humans is meaningless. Moreover, the Educational Testing Service's (ETS) own Randy E. Bennett, the Norman O. Frederiksen Chair in Assessment Innovation, has emphasized that the correlation between machine scores and the readers used for large scale test cannot be considered an accurate indicator of test validity (Bennett, 2015; Bennett & Zhang, 2015). For machine scoring to be valid, it must correlate with expert readers. The readers for state tests, however, can hardly be considered experts. Pearson, the PARCC vendor conducting the scoring, has, in the past, advertised for college graduates in any field to grade state writing tests for $12 per hour (Strauss, 2013). There was no consideration of differing educational institutions or the applicants' GPA's. In addition, Pearson readers are expected to read 20-30 essays per hour (Malady, 2013). In sum, readers are trained to grade like machines.

   The essays are extremely short, and because students have very little time, the number of words becomes the major determinant of scores both by machines and the under-qualified and over-worked human scorers (Perelman, 2012; Chodorow & Burstein, 2004). Machines are very good at counting words. The other proxy variables used are equally reductive. Development, for example, is primarily measured by the number of words and sentences in each paragraph (Quinlan, Higgins, & Wolff, 2009).

3. **In response to Question 5, asking for independent research providing evidence of the reliability of machine scoring of essays, Dr. Smith responds "The proof of concept report has three pages of references that provide evidence of reliability."** Although there are three pages of references, the 31 citations cannot in any way be considered independent research. Of these 31 citations, the lead author and most of the other authors of 28 of the references are either employees of or consultants to one of the two organizations producing the report and also employing machine scoring of essays, the Educational Testing Service and Pearson Education. Peter Foltz, one of the authors of the Proof-of-Concept Study, a Pearson Vice President, and co-developer of Pearson's Intelligent Essay Assessor scoring engine, is principal author of five of the references and co-author of five more. Jill Burstein, another one of the authors of the Proof-of-Concept Study, Research Director of the Natural Language Processing Group at ETS, and developer of ETS's e-rater scoring engine is lead author of three articles and co-author of three more. There are no references to the body of academic work, including my own, critical of the machine scoring of student essays (including Ericsson & Haswell, 2006; NCTE, 2013; Human Readers, 2013; Condon, 2013; Herrington & Moran, 2001; Herrington & Moran, 2012; Perelman, 2012; Perelman, 2013; Perelman, 2014a; Perelman, 2014b; Perelman 2016.)

4. **Other than ETS's Test of English as a Foreign Language (TOEFL), no major high stakes writing test employs machine scoring as the sole grader of student essays.** The College Board does not employ machine scoring for the SAT nor for the AP Composition Test graded by ETS. Human readers grade the ACT Essay. The other and much larger Common Core consortium of states, Smarter Balanced, is not using machines to score student English Language Arts (ELA) essays and their Proof-of-Concept Study was primarily focused on using machines as a read-behind to check the reliability of human readers. ETS has used machine scoring as a check on human readers (except for the TOEFL), although beginning this year, GRE essays will be scored by both humans and e-rater and the scores averaged to produce a final score. Moreover, some of the eight remaining PARCC states, including my own, Massachusetts, are having all essays scored by humans.

5. **The grading software is extremely susceptible to gaming.** Because the machines do not understand meaning but only count proxy variables, it is possible to attain a high score merely by providing the machine with the appropriate configuration of variables. My own research with the Basic Automatic BS Essay Language (BABEL) Generator demonstrates that the scoring engines such as e-rater will give top scores to verbose gibberish peppered with obscure language. The purpose of the exercise was to demonstrate that **if essays are graded by machines, students and teachers will have a powerful incentive to focus on bad writing, essays that are verbose and full of pretentious diction but that will not be considered gibberish by human readers and that will certainly receive top scores from the machines.**

6. **There is some indication that machine scoring can be biased towards or against certain linguistic and ethnic groups.** In an ETS research study, Ramineni et al. (2012) report that for e-rater comparisons with human readers scoring the GRE:

> Results revealed adequate performance of the different e-rater scores at the prompt level, with a notable exception for examinees from China with e-rater scores around

> half a SD higher than the human scores, and for African American test takers with e-rater scores roughly two-tenths of a SD lower than the human scores. (p. 27)

Preliminary, but not conclusive, findings in my own research on the grammar evaluation functions of machine scoring also indicate that machines would privilege English Language Learners whose first language does not possess articles, such as Mandarin or Cantonese, while over-identifying verb formation errors of English speakers whose native dialect is some form of Black English Vernacular.

Given the lack of substantive unbiased research supporting the use of machine scoring for ELA essay evaluation, the corpus of research describing various problems with it, and the issues I have outlined here, student ELA essays should be graded by human readers. Moreover, rather than having them graded by Pearson or ETS, the essays should be scored by teachers. Although more expensive, this approach has a double function of both producing much more reliable and authentic scores while simultaneously providing one of the best venues for the professional development of ELA teachers.

Les Perelman, Ph.D.

WORKS CITED

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39(1), 370-407. https://www.researchgate.net/profile/Randy_Bennett/publication/273064367_Randy_Elliot_Bennett_The_Changing_Nature_of_Educational_Assessment_Review_of_Research_in_Education_March_2015_39_370-407_doi10.31020091732X14554179/links/552ac7710cf2e089a3aa100d.pdf

Bennett, R. E., & Zhang, M. (2015). Validity and automated scoring. In F. Drasgow (Ed.), *Technology in testing: Improving educational and psychological measurement*. Washington, DC: National Council on Measurement in Education. 142-173

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, 18(1), 100-108.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater's performance on TOEFL essays (TOEFL research report, No. RR-04-73). Princeton, NJ: Educational Testing Service.

Ericsson, P. F. & Haswell, R. H. (Eds.) (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Haimson, L. (May 5, 2016) Should you trust a computer to grade your child's writing on Common Core tests? *Washington Post* https://www.washingtonpost.com/news/answer-sheet/wp/2016/05/05/should-you-trust-a-computer-to-grade-your-childs-writing-on-common-core-tests/

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English*, 63(4), 480-499.

Herrington, A., & Moran, C. (2012). Writing to a machine is not writing at all. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White*. New York, NY: Hampton Press. 219-232

Human Readers. (2013). Professionals against machine scoring of student essays in high-stakes assessment. http://humanreaders.org/petition/index.php

Lochbaum, K. E, et al. (2015). Research Results of PARCC Automated Scoring Proof of Concept Study < http://www.parcconline.org/images/Resources/Educator-resources/PARCC_AI_Research_Report.pdf>

McGraw-Hill Education CTB. (2014). Smarter Balanced Assessment Consortium Field Test: Automated Scoring Research Studies in accordance with Smarter Balanced RFP 17. < http://www.smarterapp.org/documents/FieldTest_AutomatedScoringResearchStudies.pdf>

Malady, M. J. X. (Oct. 10, 2013). We are teaching high school students to write terribly. *Slate*.

http://www.slate.com/articles/life/education/2013/10/sat_essay_section_problems_with_grading_instruction_and_prompts.html

National Council of Teachers of English (NCTE). (2013). NCTE Position Statement on Machine Scoring.
http://www.ncte.org/positions/statements/machine_scoring

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). *International advances in writing research: Cultures, places, measures*, 121-131.
http://wac.colostate.edu/books/wrab2011/chapter7.pdf

Perelman, L. (2013). Critique of Mark D. Shermis & Ben Hamner, "Contrasting State-of-the-Art Automated Scoring of Essays: Analysis." Journal of Writing Assessment 6(1)
< http://www.journalofwritingassessment.org/article.php?article=69>

Perelman, L. (2014a). When "the state of the art" is counting words. *Assessing Writing*, *21*, 104-111.

Perelman, L (2014b) Flunk the robo-graders. *Boston Globe*. April 30, 2014.
< http://www.bostonglobe.com/opinion/2014/04/30/standardized-test-robo-graders-flunk/xYxc4fJPzDr42wlK6HETpO/story.html>

Perelman, L. (2016). Grammar checkers do not work. *WLN: A Journal of Writing Center Scholarship*. 40(7-8) 11-19.
< http://lesperelman.com/wp-content/uploads/2016/05/Perelman-Grammar-Checkers-Do-Not-Work.pdf>

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). Evaluation of the e-rater® Scoring Engine for the GRE® Issue and Argument Prompts. *ETS Research Report* Series, 2012(1), i-106. < http://onlinelibrary.wiley.com/store/10.1002/j.2333-8504.2012.tb02284.x/asset/ets202284.pdf?v=1&t=ioj0n3rh&s=c57110aaca0409a1b1aa99671b9c2b01461d0ec4>

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater® scoring engine. *ETS Research Report Series*, *2009*(1), i-35.
< https://www.ets.org/Media/Research/pdf/RR-09-01.pdf>

Shermis, M D. (2014) State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing* 20, 53-76.

Shermis, M. D. (2014). The challenges of emulating human behavior in writing assessment. *Assessing Writing*, *22*, 91-99.

Strauss, V. (Jan.16, 2013). Pearson criticized for finding test essay scorers on Craigslist. *Washington Post*. < https://www.washingtonpost.com/news/answer-sheet/wp/2013/01/16/pearson-criticized-for-finding-test-essay-scorers-on-craigslist/>