Assessing Writing 21 (2014) 104–111

Contents lists available at ScienceDirect

Assessing Writing

Forum

When "the state of the art" is counting words

Les Perelman*

Massachusetts Institute of Technology, United States

ARTICLE INFO

Article history:
Received 7 May 2014
Accepted 19 May 2014
Available online 12 June 2014

Keywords:
Automated essay scoring
Common Core standard
Essay length
High-stakes assessment
Race-to-the-top
Human raters

ABSTRACT

The recent article in this journal "State-of-the-art automated essay scoring: Competition results and future directions from a United States demonstration" by Shermis ends with the claims: "Automated essay scoring appears to have developed to the point where it can consistently replicate the resolved scores of human raters in high-stakes assessment. While the average performance of vendors does not always match the performance of human raters, the results of the top two to three vendors was consistently good and occasionally exceeded human rating performance." These claims are not supported by the data in the study, while the study's raw data provide clear and irrefutable evidence that Automated Essay Scoring engines grossly and consistently over-privilege essay length in computing student writing scores. The state-of-the-art referred to in the title of the article is, largely, simply counting words.

Much of the enthusiasm for using automated essay scoring is motivated by the increased number of writing assessments informed by the Common Core standards and mandated by the U. S. Department of Education's Race-to-the-Top initiative. The stakes for getting these assessments right are very high for students, teachers, schools, school districts, and states. States are compelled by the No Child Left Behind law to use standardized test scores in teacher evaluations for tenure, pay, and promotion, as evidenced by the severe economic sanctions the Federal government has recently placed on State of Washington (Higgins, 2014). Consequently, it is inevitable that assessment will, to a large extent, define instruction. The two major Race-to-the-Top Consortia, Partnership for Assessment of Readiness for College and Careers (PARCC) and SMARTER BALANCED Assessment Consortium, are under intense

* Address: Room 14E-403, Massachusetts Institute of Technology, Cambridge, MA 02139. Tel.: +017818623833.
E-mail address: perelman@mit.edu

http://dx.doi.org/10.1016/j.asw.2014.05.001
1075-2935/

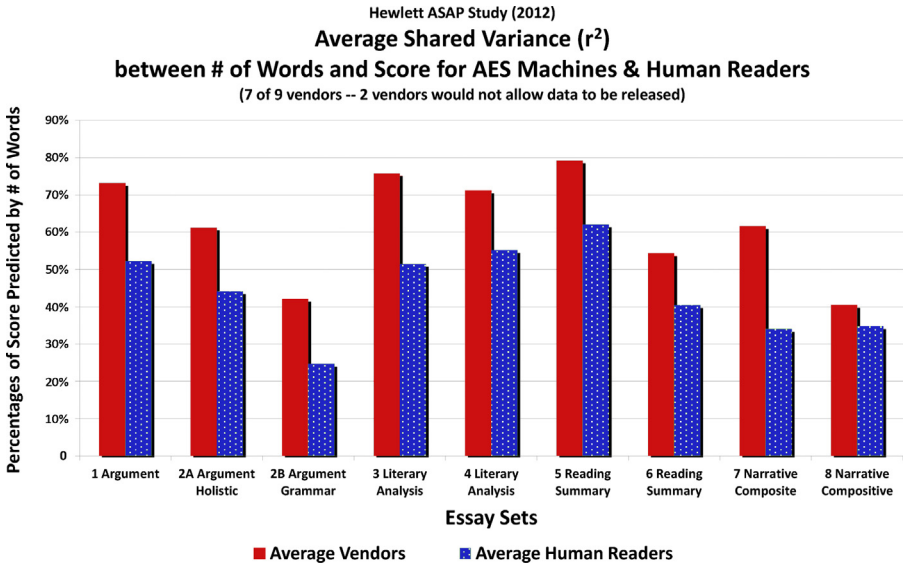

Fig. 1. Average shared variance between # of words and scores for human readers and AES machines.
Source: Calculations derived from data obtained at Automated Student Assessment Prize (2013).

pressure to cut costs. Indeed, ten of the original twenty-six PARCC states have withdrawn from the consortium largely because of cost, leaving only sixteen states and the District of Columbia (Ujifusa, 2014). At the same time, Automated Essay Scoring (AES) presents a huge economic advantage to testing companies by potentially reducing the marginal cost of scoring essays to close to zero.

It is no wonder, then, that there were large incentives to conduct the ASAP competition and to believe Professor Shermis' assertion in his article in this journal that "Automated essay scoring appears to have developed to the point where it can consistently replicate the resolved scores of human raters in high-stakes assessment" (Shermis, 2014, p. 75). Unfortunately, the data provided in that article and in the link to the raw data provided do not substantiate this claim.

The following analysis derives from three sources: the summary data presented in that article and two earlier versions of it (Shermis & Hamner, 2012, 2013), the training data downloaded from the Kaggle competition site (Kaggle, 2012), and the incomplete set of raw data from the ASAP site http://www.scoreright.org/asap.aspx?content=Request_ASAP_Phase_One_Data.

Of the nine named vendors in the study, two refused permission to have their data released. Moreover, although all participating vendors were identified in Shermis, 2014, the released raw data was anonymous, with vendors being identified only as Vendor1, Vendor2, etc. Furthermore, one of the conditions of the Terms-of-Service in downloading the data, was to refrain from any attempt to identify the participating vendors. The figure and two tables in this study are derived from my analyses of these raw data.

The principal value of Professor Shermis' study, although probably unintentional, is that the raw data of the study provide clear and irrefutable evidence that Automated Essay Scoring engines grossly and consistently over-privilege essay length in computing student writing scores. The state-of-the-art referred to in the title of the article is, in reality, simply counting words. As I have argued elsewhere (Perelman, 2012), it is this over-reliance on length that creates the apparent similarities in scores, but only for timed-impromptu writing, a genre that does not exist outside of the standardized writing test. As displayed in Fig. 1, the AES machines of the seven of nine vendors in the study that allowed their data to be released anonymously consistently overweigh word count.

The data in this figure and in both tables are reported either as correlations (the Pearson *r* product-moment correlation coefficient) or the square of the correlation, the shared variance, which is expressed as a percentage. Shared variance can be best explained as the percentage of common