Assessing Writing 22 (2014) 91–99

Contents lists available at ScienceDirect

Assessing Writing

Forum

The challenges of emulating human behavior in writing assessment

Mark D. Shermis*

Department of Educational Foundations, The University of Houston—Clear Lake, United States

ARTICLE INFO

Article history:
Available online 7 August 2014

Keywords:
Writing
Construct validity
Automated essay scoring

ABSTRACT

This is a response to Dr. Les Perelman's critique of Phase I of the Hewlett Trials. His argument is that the construct validity of the study was undermined because there was a high correlation between word count and vendor predicted scores. The response addresses the argument by showing that correlations do not mean causation. Further the reply illustrates how predications are actually formulated in automated essay scoring. The response concludes with an appeal for more research on the underlying constructs associated with writing.

1. Introduction

In his critique of my "State-of-the-art" article (2014), Dr. Perelman objected to the high correlation between raw word count and the predicted score of the automated essay scoring systems that participated in Phase I of the Hewlett Trials (Perelman, 2014). Seven of the nine vendors who participated in Phase I agreed to the release of their predicted scores under the condition that their identities be masked. Perelman's argument seems to be that this phenomenon was pervasive for the vendor-based predictions and was not particularly prevalent with the human raters; thus, the results of the study are undermined because raw word count cannot be related to the construct validity of the underlying high-stakes writing tasks.

DOI of the original article:http://dx.doi.org/10.1016/j.asw.2014.05.001.
* Tel.: +1 281 283 3501; fax: +1 281 283 3509.
E-mail address: mshermis@uhcl.edu

http://dx.doi.org/10.1016/j.asw.2014.07.002
1075-2935/

This is not a new criticism for Dr. Perelman. Almost 10 years ago, in a critique of the written portion of the SAT exam which is *graded exclusively by human raters*, he observed: "I have never found a quantifiable predictor in 25 years of grading that was anywhere near as strong as this one," he reported to the *New York Times* in 2005. "If you just graded them based on length without ever reading them, you'd be right over 90 percent of the time." According to Dr. Perelman, "the shortest essays, typically 100 words, got the lowest grade of one. The longest, about 400 words, got the top grade of six. In between, there was virtually a direct match between length and grade" (Winerip, 2005).

To bolster his argument, Dr. Perelman produced a table that contrasts the "shared variance" between the vendor predictions, the individual human rater scores, and word count for each essay set. His analysis, based on the square of the correlation coefficient (i.e., the use of *word count* as a predictive validity coefficient) shows that there are notable differences between the way human raters assign scores and the way in which automated essay scoring engines perform this task.

In addition, the critique points out where there might have been less than optimal data collection conditions. For example, the original article observed that one state appeared to make a few scoring assignments that were not consistent with their documented protocols. In the course of conducting large-scale empirical research these kinds of anomalies occur—we were fortunate to observe it and noted their existence. These and the other minor issues raised were ultimately not problematic for reviewers of the original article or they were acknowledged as limitations in the earlier study.

In the next few paragraphs I will attempt to address Dr. Perelman's main argument under the heading that *Correlation Does not Mean Causation.* I begin my response by noting the concerns that both the writing and measurement communities have expressed about the nature of the correlation, talk about why correlation does not mean causation, try to explain why in a multivariate prediction equation the bivariate relationship over-characterizes the true relationship, and suggest what the relevant communities can do to make even greater progress in getting to a point where they are talking about similar constructs underlying the evaluation of writing.

As a reminder, the primary metric against which the vendor score predictions were evaluated was a reliability coefficient called quadratic weighted kappa (κ_w) which was the relationship between the predicted essay score and the so-called "resolved score". Except for data sets 2a and 2b, the resolved score was an adjudicated resolution of the two human rater scores. In some states the score was adjudicated by adding the two human rater scores, some states took the higher of the two scores, and in some states a third rater was brought in to make a final determination of the essay score assignment. In data sets 2a and 2b, the first human rater determined the essay score assignment and the second score was only used as a check on the reliability of the first human rater. Quadratic weighted kappa is numerically equivalent to the correlation coefficient (Fleiss & Cohen, 1973). This coefficient is applied when the underlying scale has ordinal properties (e.g., higher scores indicate a better quality essay, but the distance between the scores are not necessarily equal). In practice the differences between quadratic weighted kappa and the correlation coefficient tend to be seen in the third decimal place if the prediction is constrained to an integer value which was the case in these trials. If more precise estimates are permitted, the correlation coefficient will be a higher number. Williamson, Xi, and Beyer (2012) have recognized quadratic weighted kappa as being more "rigorous" in identifying disjuncture in inter-reader reliability.

In order to pursue this discussion with a common metric, Table 1 shows the presentation of the original Pearson correlation coefficients (r) between *word count*, the *vendor predictions*, the *resolved score*, and the two *human ratings* across all nine data sets (Data Set #2 had two separate ratings). These calculations were taken from the released information that is publicly available (http://www.scoreright.org/asap.aspx?content=Request_ASAP_Phase_One_Data). Table 2 shows the same information except that all of the vendor ratings for each data set were combined and averaged. Fig. 1 plots the correlation between *word count*, the *average vendor ratings*, the *resolved score*, and the two *human ratings*. There are two relationships to note: first, the average vendor correlations with word count are higher than either the correlations with the resolved score or the human rater scores, and in some cases they differ by more than .10, a threshold recommended by Williamson et al. (2012) for flagging cases of concern (though they were examining relationships between human and automated scores, not between scores and other characteristics of essays as we are here). But the differences are not as great as Dr. Perelman's table would lead one to believe. Second, these average