Assessing Writing 34 (2017) 62–71

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

Full length article

Design and evaluation of automated writing evaluation models: Relationships with writing in naturalistic settings

Brent Bridgeman*, Chaitanya Ramineni

Educational Testing Service, United States

ARTICLE INFO

Keywords:
Automated essay scoring
Essay test validity

ABSTRACT

Automated Writing Evaluation (AWE)systems are built by extracting features from a 30 min essay and using a statistical model that weights those features to optimally predict human scores on the 30 min essays. But the goal of AWE should be to predict performance in real world naturalistic tasks, not just to predict human scores on 30 min essays. Therefore, a more meaningful way of creating the feature weights in the AWE model is to select weights that are optimized to predict the real world criterion. This unique new approach was used in a sample of 194 graduate students who supplied two examples of their writing from required graduate school coursework. Contrary to results from a prior study predicting portfolio scores, the experimental model was no more effective than the traditional model in predicting scores on actual writing done in graduate school. Importantly, when the new weights were evaluated in large samples of international students, the population subgroups that were advantaged or disadvantaged by the new weights were different from the groups advantaged/disadvantaged by the traditional weights. It is critically important for any developer of AWE models to recognize that models that are equally effective in predicting an external criterion may advantage/disadvantage different groups.

1. Introduction and literature review

Automated Writing Evaluation (AWE) typically uses computer algorithms to extract features (such as a count of grammatical errors or the number of words in a discourse element) from essays and then uses linear regression to optimally weight those features to predict a criterion. From the initial development of computer-based AWE, models were built and feature weights assigned to mimic human scores on the same essays that the machine was rating, and the quality of the machine score was assessed by its agreement with the human score in a cross-validation sample (Page, 1966). This original standard for both development and validation has been in use in the years that followed (e.g., Attali and Burstein, 2006; Dikli, 2006; Shermis & Hamner, 2013). But agreement with human scores on the same essays should not be the only validity criterion (Bennett & Bejar, 1997; Bennett & Zhang, 2016); indeed, it can be argued that such agreement is merely a reliability indicator, and not a proper validation strategy. Multiple-choice writing tests have been used as an external validity criterion with findings indicating that the multiple-choice score is predicted about equally well by human and machine essay scores (Attali, 2007). Other sub-scores from the same test have sometimes been used as an alternative validity criteria. The writing score on the Test of English as a Foreign Language (TOEFL®) is one of four scores assessing aspects of language proficiency, so the correlation of machine and human scores on the TOEFL essays with the total score from the other three sections form a reasonable, if imperfect, validation criterion. Attali (2007) and Ramineni, Trapani, Williamson, Davey, & Bridgeman (2012a) reported that human and machine scores predict this criterion about equally well. Ramineni, Trapani, Williamson,

* Corresponding author.
E-mail address: bbridgeman@ets.org (B. Bridgeman).

http://dx.doi.org/10.1016/j.asw.2017.10.001
Received 20 March 2017; Received in revised form 6 October 2017; Accepted 17 October 2017

Davey, & Bridgeman (2012b) found similar results for the Graduate Record Examinations General Test (GRE®). This test is used for admissions to graduate schools in which English is the language of instruction. It provides scores in three areas: Verbal Reasoning (V), Quantitative Reasoning (Q), and Analytical Writing (AW). The AW score is based on two thirty-minute essays; one requires the examinee to create an argument and the other requires the examinee to evaluate an argument that is presented in the prompt. Each essay is scored by a human and by computer. The research indicated that human and machine essay scores were equally effective in predicting GRE Verbal scores.

1.1. Predicting real world writing tasks

Although the ability to predict related multiple-choice scores has some value, a more rigorous test is whether automated scores from a short timed essay are related to scores on real-world writing tasks. In the United States, a substantial literature has emerged on scoring writing samples taken under naturalistic conditions (e.g.,Pullen & Haertel, 2008). These samples are often classified as performance assessments: writing tasks requiring students to construct responses over lengthy periods of time, explore alone and with others novel solutions and new task types, and reflect on those responses and solutions (Elliott et al., 2016). Bridgeman, Ramineni, Deane, & Li (2014) evaluated the ability of a machine score to predict such a naturalistic criterion. Specifically, they used portfolio scores from a sample of 352 freshman at the New Jersey Institute of Technology (NJIT). These portfolios contained:

a sampling of major writing assignments ranging from expository, memoir-style essays to a final research paper with at least one revision for assignment presented for comparison purposes. The better portfolios will attempt to organize these samplings with personal reflective statements emphasizing the specific outcomes students achieve upon their completion (personal communication from Andrew Klobucar, a faculty member at NJIT).

Portfolios were scored by NJIT faculty on a 1–6 holistic scale with an interrater reliability of 0.71. (For a detailed discussion of the development and evaluation of this portfolio assessment see Elliott, Briller, & Joshi (2007) and Elliott et al. (2016)). The portfolios were scored on the writing, reading, and critical analysis criteria elaborated in the *Framework for Success in Postsecondary Writing* (Council of Writing Program Administrators, National Council of Teachers of English, and National Writing Project, 2011). Connecting the scores to this US consensus statement lends construct validity evidence to the reported scores. In the NJIT study, human and machine scores from a two-essay placement test were the predictors. With a thirty minute time limit for each essay, students were asked to respond to two persuasive essay prompts. The correlation of the human scores with the portfolio criterion was 0.26, and the correlation for the machine scores was 0.27.

Changing the validation criterion from agreement with human raters of the same essay that the machine was evaluating to an external criterion unrelated to human scores on the initial essays was an important step, but in an even more radical break from traditional practice] Bridgeman et al. (2014, April) also argued that if the goal is to predict scores on writing samples obtained in naturalistic settings (e.g., portfolio), weights on the essay features derived from the 30 min essays could be set to optimize prediction of this criterion rather than predicting the human scores on the 30 min essays. They built such a model and observed a correlation of 0.34 (adjusted for cross-validation shrinkage) with the portfolio criterion (compared to the 0.27 when the model was built to optimize prediction human scores on the 30 min essays). They also noted that the feature weights in this experimental model differed substantially from the feature weights in the traditional model. Specifically, the most highly weighted features in the traditional model were organization basically the log of the number of discourse elements [e.g., topic and concluding sentences] and development [log of word count in a discourse element]; the combination of the organization and development feature scores is highly correlated with essay length and accounts for more than half of the relative weights in the model. In the experimental model, the organization and development features did not significantly contribute to the prediction of portfolio scores and were dropped from the model, and features related to style (e.g., overly repetitious) and word length became more important. Although essay length might be a legitimate marker for verbal fluency in a 30 min essay, it should not be surprising that length plays a diminished role in predicting portfolio scores as portfolios have no time limits.

In the present study, course-related writing samples were used for model development and evaluation. These samples were obtained as part of a comprehensive multi-university GRE predictive validity study. Developing AWE scoring models with feature weights optimized to predict scores on real-world naturalistic tasks is a new direction for machine scoring.

1.2. How a machine with no reasoning ability emulates human scores

Although a machine can realistically, though imperfectly, evaluate some essay characteristics (e.g., grammar and mechanics, word choice, and vocabulary sophistication), there are many characteristics that are clearly beyond current machine capabilities, such as providing “a cogent, well-articulated critique of the argument,” as the machine cannot distinguish a long grammatical essay with totally illogical argumentation from a long grammatical essay with appropriate argumentation. We are in full agreement with critics, such as Condon (2013) and Perelman (2014), that the current state of the art in AWE allows measurement of only a small fraction of a complete writing construct. Given these limitations, it is remarkable that the machine can replicate human scores as well as it does with human-machine agreement often as good as or better than human–human agreement (Attali, 2007; Shermis & Hamner, 2013). A plausible reason for the success of the machine is that the features that the machine can reliably score tend to be highly correlated with the argumentation skills that it cannot evaluate. There is no guarantee that a student with a sophisticated vocabulary and strong grammatical skills will necessarily create logical arguments, but in diverse samples these skills do tend occur together so that what the machine can measure is highly correlated with what it cannot assess. The rank ordering of students on the skills that the machine can evaluate tends to be highly similar to the ratings made by human scorers even though there are occasional exceptions.