Assessing Writing 22 (2014) 48–59

Contents lists available at ScienceDirect

Assessing Writing

On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): An illustration

Isaac I. Bejar*, Michael Flor, Yoko Futagi, Chaintanya Ramineni

Educational Testing Service, United States

ARTICLE INFO

Article history:
Received 18 June 2013
Received in revised form 29 May 2014
Accepted 11 June 2014
Available online 15 September 2014

Keywords:
Automated scoring
Response strategy
Validation
Essay scoring

ABSTRACT

This research is motivated by the expectation that automated scoring will play an increasingly important role in high stakes educational testing. Therefore, approaches to safeguard the validity of score interpretation under automated scoring should be investigated. This investigation illustrates one approach to study the vulnerability of a scoring engine to construct-irrelevant response strategies (CIRS) based on the substitution of more sophisticated words. That approach is illustrated and evaluated by simulating the effect of a specific strategy with real essays. The results suggest that the strategy had modest effects, although it was effective in improving the scores of a fraction of the lower-scoring essays. The broader implications of the results for quality assurance and control of automated scoring engines are discussed.

1. Introduction

As the automated scoring of constructed-response tasks becomes increasingly feasible in operation, so is the need to learn to evaluate the automated scoring engine's robustness to responses that attempt to obtain a higher score through construct-irrelevant response strategies (CIRS).[1] The purpose of this

* Corresponding author. Tel.: +1 6097345196.
E-mail address: ibejar@ets.org (I.I. Bejar).

[1] The *construct* of interest in this article is academic writing. Since we use data from a specific assessment, effectively we are relying on the corresponding definition of that construct, as described later.

http://dx.doi.org/10.1016/j.asw.2014.06.001
1075-2935/

article is to illustrate an approach to test a scoring engine's vulnerability to CIRS, and to discuss the broader implications of that threat to validity.

By all accounts, at least in the United States, there are great hopes that automated scoring will help next-generation assessments, such as those being developed in connection with the President Obama's administration Race to the Top initiative. The state consortia that have been formed to develop the primary accountability assessments are counting on the use of automated scoring. Automated scoring can lead to cost savings, but it is important to preclude the possibility that validity is reduced along the way. As Bejar (2011) discusses, quality assurance and quality control processes should be an integral part of validation for automated scoring engines. Systems can fail from time to time due to design or quality defects; therefore, there should be processes in place to detect such failures and to address them *before scores are reported*. While the results presented here are from an admissions postgraduate examination, the implications of the study are potentially applicable to other contexts, including a K-12 context, since the architecture of scoring engines for writing assessment is not assessment specific.

From a validity perspective, the vulnerability of an automated scoring engine to CIRS can erode the construct representation of scores, especially if scores were to be based solely on automated scoring. The term construct representation was introduced by Embretson (1983) to differentiate validity evidence that is internal to the test, i.e., how it is constructed, how it is scored, etc., as separate from validity evidence that is external and derives from the relationship of the scores to other variables or background variables. How a test is scored is one aspect of test design that, in conjunction with other design decisions (Bennett & Bejar, 1998), contributes to maintaining construct representation. In that sense, this study is not only about validation but also about quality assurance and control.

Specifically, the goal of this study is to evaluate the vulnerability of an automated essay scoring engine to a specific CIRS. This study is meant to be illustrative of how quality assurance checks for the automated scoring of writing can be developed, and not necessarily to develop an effective CIRS. The intention is to suggest and motivate checks in the scoring engine development process, and during the application of the scoring engine, that address the feasibility of obtaining a higher score through response strategies that do not depend on the construct of interest. Such checks could be developed proactively, or in response to strategies being used by test takers. Evidence that such checks are in place would strengthen the confidence we have in the scores.

In this article, the process of evaluating CIRS is illustrated with an earlier version (the operational version as of November 2009) of e-rater® (Attali & Burstein, 2006), by testing e-rater's vulnerability to a response strategy based on lexical alteration of the response when used to score the GRE® writing tasks.[2] The approach is simply to simulate the response strategy by altering existing essays and evaluating the impact of the strategy by comparing the scores on the altered essays to the score for the original essays. e-rater uses the frequency of the words and their length as indicators of writing quality, as well as other features. Could a simple strategy where test takers simply substitute infrequent and long words be effective in increasing scores? This is really an empirical question because there are aspects of the writing quality in an essay that are being evaluated besides lexical sophistication. Potentially, artificially increasing lexical sophistication through a substitution of words that are long and infrequent could backfire if the resulting text is evaluated less positively by the other features. That is the empirical question the study aims to answer.

2. Background for the study

Despite the improvements that have been introduced into e-rater over the years, it remains faithful to the original conception of automated scoring of essays introduced by Page (1966) several decades ago. At that time, the conceptions of validity were much narrower and emphasized prediction rather than construct validity. Additionally, the sense of "reliability" emphasized the level of inter-rater agreement (Elliot, 2005). Research at ETS in the 1960s (Godshalk, Swineford, & Coffman, 1966) established that holistic scoring was a feasible approach to improve inter-rater agreement that could be implemented on a large scale. Holistic scoring was economical to boot, and thus, it became the standard

[2] This study used data from the previous edition of the GRE. A new revised GRE was launched in August, 2011.