Assessing Writing 20 (2014) 53–76

Contents lists available at ScienceDirect

Assessing Writing

State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration

Mark D. Shermis [a,b,*]

[a] Department of Educational Foundations and Leadership, The University of Akron, United States
[b] Department of Psychology, The University of Akron, United States

ARTICLE INFO

Article history:
Received 27 August 2012
Received in revised form 17 April 2013
Accepted 26 April 2013
Available online 30 January 2014

Keywords:
Automated essay scoring
High-stakes assessment
Writing
Race-to-the-Top
Performance assessment
Human raters

ABSTRACT

This article summarizes the highlights of two studies: a national demonstration that contrasted commercial vendors' performance on automated essay scoring (AES) with that of human raters; and an international competition to match or exceed commercial vendor performance benchmarks. In these studies, the automated essay scoring engines performed well on five of seven measures and approximated human rater performance on the other two. With additional validity studies, it appears that automated essay scoring holds the potential to play a viable role in high-stakes writing assessments.

1. Introduction

1.1. Context

A new generation of measurement instruments is being planned for use in the United States as part of the Race-to-the-Top assessments. These instruments will be based on the instructional goals of Common Core State Standards that articulate the required proficiencies for United States students

* Correspondence to: Department of Educational Foundations and Leadership, The University of Akron, 213 Crouse Hall, 44325, United States. Tel.: +1 954 899 8069.
E-mail address: mshermis@uakron.edu

1075-2935/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.asw.2013.04.001

to be "college-ready" by the time they graduate from high school (Porter, McMaken, Hwang, & Yang, 2011). These new assessments will likely rely less on multiple-choice questions and use performance measures that more closely match the construct under investigation. The move to Common Core State Standards in the U.S. represents a significant departure from a curricular structure that heretofore has been driven by individual states.

Over the past 30 years the high-stakes assessments associated with state objectives have been calibrated to the minimal standards for exiting high school. These standards have not been universal and vary from state to state. In the area of high-stakes state writing assessment, writing objectives can range from the summarization of reading material to the ability to create prose of a particular genre to the mastery of a particular writing form. Writing assessment practices also differ from state to state including the amount and type of writing expected, types of rubrics used, scoring and adjudication protocols, the number and qualifications of raters employed, quality assurance practices, and the reporting of results.

In part because of the emphasis on minimum competency and the varied nature of what a state might emphasize in their high-stakes testing programs there grew a widening pool of college students who had the skill set to graduate from high school yet had to enroll in remedial college classes because that skill set did not include the higher order knowledge or skills required to perform well in entry-level college classes (Attewell, Lavin, Domina, & Levey, 2006) where the curriculum is typically based on the standards of the discipline's national organization. The two major Race-to-the-Top Consortia [Partnership for Assessment of Readiness for College and Careers (PARCC) and SMARTER Balanced Assessment Consortia (SBAC)] and their 46 subscriber states intend to change that pattern by having all students – even those who may wish to pursue a vocational track – work toward college readiness rather than a mastery of basic high school skills (Tucker, 2009).

With regard to assessments in English Language Arts and in many of the science areas, this shift will mean more writing. For instance, students might be given an array of articles in biology to read and then respond to an essay prompt that addresses some conclusion that they might make based on the articles. The essay might ask the student to explain a rationale for a conclusion or to cite evidence in support of an argument. Part of the current debate in planning the new instruments is whether this performance assessment is really a writing task (where the emphasis is on writing ability), reading comprehension (where the emphasis is on understanding the content), or one of critical thinking (where the emphasis is on synthesizing and evaluating information). Two of these options are consistent with Weigle's distinctions regarding the multiple purposes of assessment, assessing writing (AW) and assessing content through writing (ACW) (Weigle, 2013; Weir, 2005). In many cases, the student will be asked to produce a written artifact that must be evaluated – and to do so numerous times throughout the academic year. The sheer number of written responses for high-stakes summative assessments across the grade levels makes it challenging and cost-ineffective to have human raters exclusively score these assessments. For example, the state of Florida has approximately 180,000 students in each grade level. If each student in that one state had five essays graded, the state would be required to evaluate almost 11 million documents per year, raising questions as to the feasibility of recruiting a sufficient number of qualified human graders to provide final scores, read reliably, in a timely manner across the entirety of the United States. The goals of the Consortia have been to strongly encourage the development and use of machine scoring algorithms in order to make it possible to score such volumes in a timely and cost-effective manner.

In order to evaluate the basic feasibility of these goals, the Hewlett Foundation (www.hewlett.org) sponsored a demonstration of existing and emerging automated scoring systems for essays as part of the Automated Student Assessment Prize (ASAP) program (Shermis & Hamner, 2012, 2013), the results of which are reported here. ASAP is an independently funded organization that is exploring the effectiveness of machine scoring in different contexts and sponsors open and public prize competitions to stimulate innovations in machine scoring. Two studies are described below; one is a demonstration of performance for existing essay scoring engines, and another an open competition designed to encourage the creation of new algorithms to score essays. For both studies the goal was to evaluate the extent to which automated scoring systems for essays are capable of producing scores similar to those of trained human graders. The first study focused on pre-existing systems for automated scoring of essays and compared the systems of eight commercial vendors and one university laboratory