An Overview of Automated Scoring of Essays
Samire Dikli

Abstract

Automated Essay Scoring (AES) is defined as the computer technology that evaluates and scores the written prose (Shermis & Barrera, 2002; Shermis & Burstein, 2003; Shermis, Raymat, & Barrera, 2003). AES systems are mainly used to overcome time, cost, reliability, and generalizability issues in writing assessment (Burstein, 2003; Burstein, 2003; Chung & O’Neil, 1997; Hamp-Lyons, 2001; Myers, 2003; Page, 2003; Rudner & Gagne, 2001; Rudner & Liang, 2002; Sireci & Rizavi, 1999; http://people.emich.edu). AES continue attracting the attention of public schools, universities, testing companies, researchers and educators (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Shermis & Burstein, 2003; Sireci & Rizavi, 1999). The main purpose of this article is to provide an overview of current approaches to AES. After describing the most widely used AES systems (i.e., Project Essay Grader (PEG), Intelligent Essay Assessor (IEA), E-rater and Criterion, IntelliMetric and MY Access, and Bayesian Essay Test Scoring System (BETSY)), main characteristics of these systems will be discussed and current issues regarding the use of them both in low-stakes assessment (in classrooms) and high-stakes assessment (as standardized tests) will be discussed in this article.

Keywords

Computerized writing; assessment; technology; writing; automated; scoring; essays; AES; essay; computer; automated scoring

Full Text:

Automated Essay Scoring (AES) is defined as the computer technology that evaluates and scores the written prose (Shermis & Barrera, 2002; Shermis & Burstein, 2003; Shermis, Raymat, & Barrera, 2003). The main purpose of this article is to provide an overview of current approaches to AES. After describing the most widely used AES systems (i.e., Project Essay Grader (PEG), Intelligent Essay Assessor (IEA), E-rater and Criterion, IntelliMetric and MY Access, and Bayesian Essay Test Scoring System (BETSY)), main characteristics of these systems will be discussed and current issues regarding the use of them both in low-stakes assessment (in classrooms) and high-stakes assessment (as standardized tests) will be discussed in this article. Keywords. The use of automated essay scoring is also somewhat controversial. A well-written essay about baking a cake could receive a high score if PEG were used to grade essays about causes of the American Civil War. Conceivably, IEA could be tricked into giving a high score to an essay that was a string of relevant words with no sentence structure whatsoever. E-rater appears to overcome some of these criticisms at the expense of being fairly complicated. One should not expect perfect accuracy from any automated scoring approaches. The correlation of human ratings on state assessment constructed-response items is typically only .70 -.75. Thus, correlating with human raters as well as human raters correlate with each other is not a very high, nor very meaningful, standard.