Assessing Writing 18 (2013) 25–39

Contents lists available at SciVerse ScienceDirect

Assessing Writing

Automated essay scoring: Psychometric guidelines and practices

Chaitanya Ramineni*, David M. Williamson

Educational Testing Service, Rosedale Rd, Princeton 08541, NJ, USA

ARTICLE INFO

Article history:
Available online 13 November 2012

Keywords:
Automated essay scoring
Constructed response items
Large-scale writing assessments

ABSTRACT

In this paper, we provide an overview of psychometric procedures and guidelines Educational Testing Service (ETS) uses to evaluate automated essay scoring for operational use. We briefly describe the e-rater system, the procedures and criteria used to evaluate e-rater, implications for a range of potential uses of e-rater, and directions for future research. The description of e-rater includes a summary of characteristics of writing covered by e-rater, variations in modeling techniques available, and the regression-based model building procedure. The evaluation procedures cover multiple criteria, including association with human scores, distributional differences, subgroup differences and association with external variables of interest. Expected levels of performance for each evaluation are provided. We conclude that the *a priori* establishment of performance expectations and the evaluation of performance of e-rater against these expectations help to ensure that automated scoring provides a positive contribution to the large-scale assessment of writing. We call for continuing transparency in the design of automated scoring systems and clear and consistent expectations of performance of automated scoring before using such systems operationally.

1. Introduction

There are many areas of practice, both educational and professional, in which the ability of interest is not based solely on knowledge but rather includes an expectation of performance as well. Writing is one such ability. As such, proficiency in writing calls for more than the ability to *recognize* distinctions

* Corresponding author at: MS 18-E, ETS 660 Rosedale Rd, Princeton 08541, NJ, USA. Tel.: +1 609 734 5403.
E-mail address: cramineni@ets.org (C. Ramineni).

1075-2935/$ – see front matter © 2012 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.asw.2012.10.004

between higher and lower quality of writing, but also to *produce* writing that is of a certain quality. As a result, assessments of writing ability routinely include more than multiple-choice items and incorporate one or more tasks that require the production of writing.

Tasks requiring the production of a response rather than the selection of a response are referred to as *constructed-response items*. Inclusion of constructed-response items in assessment can provide for greater construct representation (greater fidelity to the abilities of interest) but pose challenges in the reliability of scoring, timeliness of reporting scores, and the time required for test administration. Despite these challenges, constructed-response writing tasks have been added to several testing programs in recent years, including the Graduate Record Examination (GRE®), the Test of English as a Foreign Language iBT (TOEFL®), and the SAT®, among others.

While inclusion of writing tasks expands the representation of the writing construct, it poses challenges for high-quality scoring of large volumes of essays in a timely manner. Automated essay scoring (AES) systems hold the potential for greater use of essays in assessment while also maintaining the reliability of scoring and the timeliness of score reporting desired for large-scale assessment. While this potential is appealing, we need to know when AES systems are of sufficient quality to be relied upon for scoring, particularly when the assessment can have important consequences for individuals, such as in college admissions. In this paper we offer some perspective on evaluation procedures and criteria used by Educational Testing Service (ETS) for determining whether the e-rater® AES should be part of scoring an assessment (Williamson, Xi, & Breyer, 2012). We begin with a brief orientation to the contexts of use for constructed-response items and automated scoring, broadly defined. We then provide some background on available AES systems and a sampling of applications, transitioning into an orientation to the e-rater AES system that is the focus of the paper. The core of this paper is focused on the evaluation procedures applied to e-rater and expectations of performance for operational use, with implications for how these impact decision-making regarding use of AES in assessment.

2. Background

There is a natural tension between the demands of practice, the naturalistic behaviors that occur when participating in some activity, and the demands of assessment, a set of predefined circumstances intended to evaluate ability. In naturalistic environments the tasks undertaken may be very "rich" and represent complex, labor-intensive efforts carried out over extended periods. However, they may also be highly idiosyncratic and variable based on the particular circumstances, timing and characteristics of the task undertaken. By contrast, assessment often emphasizes control of the circumstances to elicit certain behaviors and standardization of the required tasks so that fair comparisons can be made among examinees. Control and standardization can come at the expense of perceived fidelity to the real life behaviors that the assessment tasks are designed to represent. Constructed-response items, in which an examinee produces a response in a more naturalistic manner rather than selecting a response from a set of pre-defined options (as in multiple-choice items), represent an effort to provide some connection between the naturalistic characteristics of real world performance and the controlled circumstances of assessment. In so doing, a certain degree of real-world fidelity is sacrificed for the sake of assurances of individual effort, fair comparisons across individuals, and unambiguous scoring while at the same time some test development goals, such as maximizing reliability, efficient use of test time, cost-effectiveness and fast score reporting are compromised to provide for greater fidelity to practice.

There are a number of domains of practice that are sufficiently performance oriented that consequential assessments include constructed-response items. Examples include: use of computer-aided design in architectural design tasks (Braun, Bejar, & Williamson, 2006); use of standardized patients in physician licensure (Margolis & Clauser, 2006); accounting (DeVore, 2002); state educational achievement tests (The Florida Comprehensive Assessment Test); and, of course, writing as part of large scale admissions testing (Breland, Kubota, Nickerson, Trapani, & Walker, 2004; Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012a; Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012b). However, the incorporation of constructed-response tasks results in additional challenges for assessment. When compared to their multiple-choice counterparts, constructed-response items take longer to administer and provide less psychometric information per unit time of assessment. They also reduce