Automated Essay Scoring with the E-rater System

Yigal Attali
Educational Testing Service
yattali@ets.org

Abstract

This paper provides an overview of *e-rater*®, a state-of-the-art automated essay scoring system developed at the Educational Testing Service (ETS). E-rater is used as part of the operational scoring of two high-stakes graduate admissions programs: the *GRE*® General Test and the *TOEFL iBT*® assessments. E-rater is also used to provide score reporting and diagnostic feedback in *Criterion*^{SM}, ETS's writing instruction application.

E-rater scoring is based on automatically extracting features of the essay text using Natural Language Processing (NLP) techniques. These features measure several underlying aspects of the writing construct: word choice, grammatical conventions (grammar, usage, and mechanics), development and organization, and topical vocabulary usage.

The paper reviews e-rater's feature set, the framework for feature aggregation into essay scores, processes for the evaluation of e-rater scores, and options for operational implementation, with an emphasis on the standards and procedures used at ETS to ensure scoring quality.

Keywords: automated scoring, writing assessment

The use of essay writing assessments in large-scale testing programs has been greatly expanded in recent years, including the SAT, GRE, TOEFL, and GMAT testing programs, to name a few prominent examples. These assessments usually consist of one or two writing tasks. The tasks are timed (typically with a time limit of 25 to 45 minutes) and consist of a topic (or prompt) the student is asked to write about. For example, the Analytical Writing measure of the GRE Revised General Test comprises two essay writing tasks. In the issue task, the student is asked to discuss and express his or her perspective on a topic of general interest. In the argument task, a brief passage is presented in which the author makes a case for some course of action or interpretation of events by presenting claims backed by reasons and evidence. The student's task is to discuss the logical soundness of the author's case by critically examining the line of reasoning and the use of evidence.

Human evaluation of essay responses is based on a rubric that delineates specific expectations about essay responses. The rubric describes (typically 4-6) levels of performance across different writing quality dimensions such as development and organization, word choice, sentence fluency, and conventions. Graders may be expected to form a "holistic" impression of the essay response or evaluate it along several dimensions.

As measures of writing skill, essay writing assessments are often favored over measures that assess students' knowledge of writing conventions (for example, through multiple-choice tests), because they require students to produce a sample of writing and as such are more "direct." However, a drawback of large-scale essay writing assessments is that their evaluation is a complex and time-consuming process associated with significant costs. These difficulties have led to a growing interest in the application of automated natural language processing techniques for the development of automated essay scoring as an alternative or complement to human scoring of essays.

Automated essay scoring (AES) technologies have a relatively long history (Page, 1966) and several commercial applications exist today. This paper describes the e-rater system (Attali & Burstein, 2006; Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Burstein, Tetreault, & Madnani, 2013), initially developed by Educational Testing Service in the 1990s. E-rater is used as part of the operational scoring of two high-stakes graduate admissions programs: the GRE General Test and the TOEFL iBT assessments. E-rater is also used to provide score reporting and diagnostic feedback in *Criterion*^{SM}, ETS's writing instruction application.

Scoring

E-rater features

AES systems do not actually read and understand essays as humans do. Whereas human raters may directly evaluate various intrinsic variables of interest, such as diction, fluency, and grammar, in order to produce an essay score, AES systems use approximations or possible correlates of these intrinsic variables.

The e-rater architecture is based on a small set of measures (or features) that were developed to cover different aspects of the writing construct. These features may themselves be based on numerous micro-features and underlying systems. In addition, the features can be clustered into four groups of features that cover four facets of the writing construct: word choice, grammatical conventions, fluency and organization, and topical vocabulary usage.

The word choice facet is measured by two features. The first is a measure of vocabulary level based on relative occurrence of words in written texts. The second feature is based on the average word length in characters across the words in the essay.

The grammatical conventions facet is measured by four features. A grammar feature is based on rates of errors such as fragments, run-on sentences, garbled sentences, subject-verb agreement errors, ill-formed verbs, pronoun errors, missing possessives, and wrong or missing words. A usage feature is based on rates of errors such as wrong or missing articles, confused words, wrong form of words, faulty comparisons, and preposition errors. A mechanics feature is based on rates of spelling, capitalization, and punctuation errors. A fourth feature evaluates correct usage of collocations (e.g., “powerful computer” versus “strong computer”) and prepositions.

The fluency and organization facet is measured by four features. The first is a measure of organization based on detection of discourse elements (i.e., introduction, thesis, main points, supporting ideas, and conclusion) in the text. A development feature is based on the relative development of these discourse elements. A style feature is based on rates of cases such as overly repetitious words, inappropriate use of words and phrases, sentences beginning with coordinated conjunctions, very long and short sentences, and passive voice sentences. A sentence variety feature is based on evaluating the range and quality of syntactic patterns in the text.

Finally, the topical vocabulary usage facet is measured by features that compare the vocabulary of the essay with typical vocabulary found in high- and low-quality essays written on the same topic or on other topics of the same assessment task.

In addition to essay scoring features, e-rater also includes systems that are designed to identify anomalous and off-topic essays. Such essays are flagged and not scored by e-rater.

Feature aggregation

In order to report an essay score, the feature scores need to be aggregated. Essay scoring in e-rater is a relatively straightforward process. E-rater scores are calculated as a weighted average of the feature values (after appropriate feature standardization is applied), followed by applying a linear transformation to achieve a desired scale.

A major issue in developing a scoring scheme is the determination of feature weights. Weights represent the relative importance of the different features. Different weighting schemes will result in scores that have different meanings depending on the features that are emphasized. Therefore, the choice of a weighting scheme and its rationale is of utmost importance for the validity of AES. Traditionally, weighting schemes for AES have been based almost exclusively on the concept of optimizing the relation of automated scores with human scores of the same essays. For example, by using multiple regression to predict human scores from the set of features calculated on the same essays, weights will be obtained that maximize the relation between the predicted scores and the human scores on this set of essays. Although e-rater scoring is usually based on this approach, other alternatives have also been considered and are under research. One such alternative is judgment-based weighting, where experts set weights according to judged importance of AES features or measured dimensions of writing (Ben-Simon & Bennett, 2007). Another alternative is a factor-analytic approach, where weights are based on the internal structure that underlies the measurement of features (Attali, 2012).

Another major issue in scoring is the appropriate level of the scoring model. Traditionally, AES systems are trained and calibrated separately for each prompt (e.g., Landauer, Laham, & Foltz, 2003; Page, 1994). This means that the features used, their weights, and scoring standards, may be different across prompts of the same assessment. Consequently, scores will have different meanings across prompts. With e-rater, the small and standardized feature set allows for the possibility of applying the same scoring standards across all prompts of an assessment. For example, the effect of a particular grammar score on the essay score would be the same across prompts. Such a “generic” scoring approach produces standardized scores across prompts, and is more consistent with the human rubric that is usually the same for all assessment prompts, and thus contributes to the validity of scores. It also offers substantive logistical advantages for large-scale assessments because it allows scoring essays from new prompts without first training the system on each specific prompt. E-rater routinely applies a single scoring model across all prompts of a writing assessment (Attali, Bridgeman, & Trapani, 2010). Moreover, Attali and Powers (2009) extended the notion of the generic model across assessments and ability levels, by creating a developmental writing scale based on e-rater features, a single scoring model (and standards) for timed writing performance of children from 4th to 12th grade.

A third important issue in scoring is the consideration of sub-scores, or trait scores, in addition to overall essay scores. Trait scores can capture examinees’ specific weaknesses and strengths in writing, but have often proven less useful than expected because they are highly correlated among themselves and with holistic scores, thus rendering them redundant from a psychometric point of view. With e-rater, research (Attali, 2011) has shown that trait scores based on the four major facets measured by e-rater features (word choice, grammatical conventions, fluency/organization, and topical vocabulary) are sufficiently reliable and independent to provide added psychometric value beyond the overall essay score.

Performance evaluation

The quality of e-rater scores has been subjected to numerous evaluations. This section provides an overview of the types of evaluations that are performed to ensure the technical soundness of an e-rater implementation.

Construct relevance

An initial step in the use of e-rater for a particular assessment is an evaluation of its fit with the goals, design, and scoring rubric of the assessment (Williamson, 2009). The GRE issue task and the TOEFL independent task both ask the examinee to support an opinion on a provided topic, and elicit relatively less constrained responses from examinees. On the other hand, the GRE argument task asks examinees to critique and discuss the reasoning of a given argument, and the TOEFL integrated task requires examinees to read a passage, listen to a related lecture, and finally write in response to what they have read and heard. These tasks elicit relatively more constrained responses, and their evaluation by human raters is more intimately related to content of the prompt.

Consequently, e-rater may not be equally capable of measuring the construct intended by these two types of tasks.

Reliability

Reliability is concerned with consistency of test scores and is based on the idea that the observed score on a test is only one possible result that might have been obtained under different conditions – another occasion of the test or a different form of the test. The reliability of e-rater was estimated in several studies. For example, Attali and Powers (2009) estimated reliability coefficients of .67 to .75 for scores of 4th to 12th grade students who participated in a research study and submitted four essays within a timeframe of a few weeks. Attali (2012) reports cross task reliability coefficients for GRE and TOEFL, which can be thought of as lower bounds for same-task reliability estimates. The correlation between GRE argument and issue e-rater scores was .75 and the correlation between TOEFL independent and integrated e-rater scores was .70. All the above figures estimate the reliability of a single essay test. It should be noted that the reported reliability of human scores is considerably lower. For example, the cross task reliability coefficients for a single human rater in Attali (2012) are .56 and .51 for GRE and TOEFL, respectively (about .20 lower than e-rater reliability).

Association with human scores

AES evaluations have traditionally used agreement between automated and human scores for the same essay responses as the main performance criterion. For e-rater, correlations between a human rating and e-rater scores have been found to be similar to correlations between two human ratings. For example, Ramineni, Trapani, Williamson, Davey, & Bridgeman (2012a) show that for the GRE issue task human-human correlations (.74) are lower than human-machine correlations (.80), but for the GRE argument task human-human correlations (.78) are at least as high as human-machine correlations (.78). Similarly, Ramineni, Trapani, Williamson, Davey, & Bridgeman (2012b) show that for the TOEFL independent task human-human correlations (.69) are lower than human-machine correlations (.75), but for the TOEFL integrated task human-human correlations (.82) are higher than human-machine correlations (.73). These differences can be explained by considering the writing requirements of the four tasks (see above).

Subgroup differences

Since writing is a complex construct, groups of examinees with different linguistic and cultural backgrounds could develop distinct patterns of writing skills. Because machine scores measure the writing construct differently than human scores, these skills may be more or less influential in human or machine scores. Consequently, specific groups of examinees may have, on average, higher or lower machine scores than human scores. Bridgeman, Trapani, and Attali (2012) explored this possibility across gender, ethnic, and country of origin groups, for both the GRE and TOEFL. Human and machine scores were very similar across most subgroups, but there were some notable exceptions. Chinese speaking examinees, and in particular examinees from mainland China, earned relatively higher e-rater scores than human scores, and Arabic speaking examinees earned relatively lower e-rater scores. These results were in the same direction as found by Burstein and Chodorow (1999) with an earlier version of e-rater. Although these results could emerge from distinct writing styles of certain language groups, lack of differences for countries that share the same or similar languages (Korea, Japan, and even Chinese-speaking Taiwan) suggest that cultural differences, or even test preparation practices, could also account for these findings.

Relation with other measures

Although human scores on the same essays are used as the main performance criterion for automated scores, investigations of the relationships of human and automated scores with other measures of the same or a similar construct could serve as convergent evidence against the threat of construct under-representation. For example, several studies examined the relations of e-rater and human essay scores with other sub-scores from the same general ability test (Attali, 2007; Attali, 2009; Ramineni et al., 2012a, 2012b). These measures are relevant for validation of essay scores especially when their purpose is to measure other aspects of language proficiency (like reading, speaking and listening for TOEFL and verbal scores for GRE), because they can serve as additional convergent evidence on the validity of scores. The general finding in the above studies is that human and machine essay scores correlate roughly equally with other sub-scores.

The relationship with measures of other proficiencies, such as mathematical reasoning, can serve as further evidence that machine scores are not influenced by construct irrelevant factors. Ramineni et al. (2012a) found that GRE quantitative scores had similar correlations with e-rater (.13 and .24 for argument and issue) and with human scores (.07 and .22).

Consequences of use

An evaluation of the intended and unintended consequences of AES use should be part of the evidence for the validity of automated scores. A possible unintended consequence of AES use is that students change their writing strategies to accommodate automated scoring. Although evidence for this possibility can indirectly be drawn from monitoring of human and machine scores (as well as individual features), a useful way to approach this issue is to ask test takers directly about their reactions to AES and whether they had approached test taking differently as a result of the use of automated scoring. Powers (2011) surveyed TOEFL test takers and asked them what would be good strategies for writing essays that are scored automatically. The most frequently endorsed strategies were to pay more attention to spelling and grammar and to the form or structure of the essay, such as making certain to include sections like an introduction and a conclusion. Other popular strategies were to use more transition words and more diverse vocabulary. On the other hand, test takers were far less likely to endorse such strategies as writing lengthy essays or using complex sentences and long words. In addition, most did not think it a good strategy to focus less on the content or logic of their essays.

Variations in Use

E-rater scores are employed in different ways, depending on their intended use and the stakes associated with them. In *Criterion*, ETS's writing instruction application, students receive instantaneous evaluation and diagnostic feedback on their writing. This evaluation is based on the e-rater engine and is primarily used by teachers as a practice tool and instructional aid.

Because GRE and TOEFL scores have much higher stakes, e-rater scores are combined with human evaluation in these programs. For the TOEFL tasks, a single human rating is averaged with the e-rater score. For one of the tasks, the human and machine scores have equal weights in calculating this average, whereas in the other task the human score is weighted twice as the e-rater score. In GRE, e-rater scores do not contribute directly to the final score. Instead, they are used as a quality control (a “check”) on human ratings in the following way. The results of a single human rating and the automated score are compared. If there is a discrepancy beyond a certain threshold between the two, the essay is scored by a second human rater. In both cases, the reported score is based only on the human scores, either the single human score or the average of the two human scores. This is the approach implemented for the GRE issue and argument tasks.

References

Attali, Y. (2011). *Automated subscores for TOEFL iBT Independent essays* (ETS RR-11-39). Educational Testing Service: Princeton, NJ.

Attali, Y. (2012, April). *Factor structure of the e-rater automated essay scoring system*. Paper presented at the annual meeting of the National Council of Measurement in Education, Vancouver, CA.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-198). New York, NY: Routledge.

Attali, Y., Bridgeman, B., & Trapani, C. S. (2010). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning, and Assessment, 10*(3). Available from http://www.jtla.org.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Available from http://www.jtla.org .

Attali, Y., & Powers, D. (2009). Validity of scores for a developmental writing scale based on automated scoring. *Educational and Psychological Measurement, 69*, 978-993.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87-112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ben-Simon, A. & Bennett, R. E. (2007). Toward More Substantively Meaningful Automated Essay Scoring. *Journal of Technology, Learning, and Assessment, 6*(1). Retrieved 3/9/2012 from http://www.jtla.org.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In M. Broman Olsen (Ed.), *Computer mediated language assessment and evaluation in natural language processings* (pp. 68–75). Morristown, NJ: Association for Computational Linguistics.

Burstein, J., Kukich, K.,Wolff, S., Lu, C., & Chodorow, M. (1998, April). *Computer analysis of essays*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55-67). New York, NY: Routledge.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87-112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238-243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*, 127-142.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012a). *Evaluation of e-rater® for the GRE® issue and argument prompts* (ETS RR-12-02). Educational Testing Service: Princeton, NJ.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012b). *Evaluation of e-rater® for the TOEFL® independent and integrated prompts* (ETS RR-12-03). Educational Testing Service: Princeton, NJ.

Williamson, D. M. (2009, April). *A framework for implementing automated scoring*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.