Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 7, Number 26, December, 2001 ISSN=1531-7714

An Overview of Three Approaches to Scoring Written Essays by Computer

Lawrence Rudner and *Phill Gagne*
University of Maryland, College Park

It is not surprising that extended-response items, typically short essays, are now an integral part of most large-scale assessments. Extended response items provide an opportunity for students to demonstrate a wide range of skills and knowledge, including higher order thinking skills such as synthesis and analysis. Yet assessing students' writing is one of the most expensive and time-consuming activities for assessment programs. Prompts need to be designed, rubrics created, multiple raters need to be trained, and then the extended responses need to be scored, typically by multiple raters. With different people evaluating different essays, interrater reliability becomes an additional concern in the writing assessment process. Even with rigorous training, differences in the background, training, and experience of the raters can lead to subtle but important differences in grading (Blok & de Glopper, 1992, Rudner, 1992).

Computers and artificial intelligence have been proposed as tools to facilitate the evaluation of student essays. In theory, computer scoring can be faster, reduce costs, increase accuracy and eliminate concerns about rater consistency and fatigue. Further, the computer can quickly rescore materials should the scoring rubric be redefined. This articles describes the three most prominent approaches to essay scoring.

Systems

The most prominent writing assessment programs are

- Project Essay Grade (PEG), introduced by Ellis Page in 1966,
- Intelligent Essay Assessor (IEA), first introduced for essay grading in 1997 by Thomas Landauer and Peter Foltz, and
- E-rater, used by Educational Testing Service (ETS) and developed by Jill Burstein.

Descriptions of these approaches can be found at the web sites listed at the end of this article and in Whittington and Hunt (1999) and Wresch (1993). Other software projects are briefly mentioned in Breland and Lytle (1990), Vetterli and Furedy (1997), and Whissel (1994).

Page uses a regression model with surface features of the text (document length, word length, and punctuation) as the independent variables and the essay score as the dependent variable. Landauer's approach is a factor-analytic model of word co-occurrences which emphasizes essay content. Burstein uses a regression model with content features as the independent variables.

PEG - PEG grades essays predominantly on the basis of writing quality (Page, 1966, 1994). The underlying theory is that there are intrinsic qualities to a person's writing style called trins that need to be measured, analogous to true scores in measurement theory. PEG uses approximations of these variables, called proxes, to measure these underlying traits. Specific attributes of writing style, such as average word length, number of semicolons, and word rarity are examples of proxes that can be measured directly by PEG to generate a grade. For a given sample of essays, human raters grade a large number of essays (100 to 400), and determine values for up to 30 proxes. The grades are then entered as the criterion variable in a regression equation with all of the proxes as predictors, and beta weights are computed for each predictor. For the remaining unscored essays, the values of the proxes are found, and those values are then weighted by the betas from the initial analysis to calculate a score for the essay.

Page has over 30 years of research consistently showing exceptionally high correlations. In one study, Page (1994) analyzed samples of 495 and 599 senior essays from the 1998 and 1990 National Assessment of Educational Progress using responses to a question about a recreation opportunity: whether a city government should spend its recreation money fixing up some abandoned railroad tracks or converting an old warehouse to new uses. With 20 variables, PEG reached multiple Rs as high as .87, close to the apparent reliability of the targeted judge groups.

IEA - First patented in 1989, IEA was designed for indexing documents for information retrieval. The underlying idea is to identify which of several calibration documents are most similar to the new document based on the most specific (i.e., least frequent) index terms. For essays, the average grade on the most similar calibration documents is assigned as the

computer-generated score (Landauer, Foltz, Laham, 1998).

With IEA, each calibration document is arranged as a column in a matrix. A list of every relevant content term, defined as a word, sentence, or paragraph, that appears in any of the calibration documents is compiled, and these terms become the matrix rows. The value in a given cell of the matrix is an interaction between the presence of the term in the source and the weight assigned to that term. Terms not present in a source are assigned a cell value of 0 for that column. If a term is present, then the term may be weighted in a variety of ways, including a 1 to indicate that it is present, a tally of the number of times the term appears in the source, or some other weight criterion representative of the importance of the term to the document in which it appears or to the content domain overall.

Each essay to be graded is converted into a column vector, with the essay representing a new source with cell values based on the terms (rows) from the original matrix. A similarity score is then calculated for the essay column vector relative to each column of the rubric matrix. The essay's grade is determined by averaging the similarity scores from a predetermined number of sources with which it is most similar. Their system also provides a great deal of diagnostic and evaluative feedback. As with PEG, Foltz, Kintsch and Landauer (1998) also report high correlations between IEA scores and human scored essays.

E-rater - The Educational Testing Service's Electronic Essay Rater (e-rater) is a sophisticated "Hybrid Feature Technology" that uses syntactic variety, discourse structure (like PEG) and content analysis (like IEA). To measure syntactic variety, e-rater counts the number of complement, subordinate, infinitive, and relative clause and occurrences of modal verbs (would, could) to calculate ratios of these syntactic features per sentence and per essay. For structure analysis, e-rater uses 60 different features, similar to PEG's proxes.

Two indices are created to evaluate the similarity of the target essay's content to the content of calibrated essays. As described by Burstein, et.al (1998), in their *EssayContent* analysis module, the vocabulary of each score category is converted to a single vector whose elements represent the total frequency of each word in the training essays for that holistic score category. The system computes correlations between the vector for a given test essay and the vectors representing the trained categories. The score that is most similar to the test essay is assigned as the evaluation of its content. E-rater's *ArgContent* analysis module is based on the inverse document frequency, like IEA. The word frequency vectors for the score categories are converted to vectors of word weights. Scores on the different components are weighted using regression to predict human grader's scores.

Analysis

Several studies have reported favorably on PEG, IEA, and e-rater. A review of the research on IEA found that its scores typically correlate as well with human raters as the raters do with each other (Chung & O'Neil, 1997). Research on PEG consistently reports relatively high correlations between PEG and human graders relative to correlations between human graders (e.g., Page, Poggio, & Keith, 1997). E-rater was deemed so impressive it is now operational and used to score the General Management Aptitude Test (GMAT). All of the systems return grades that correlate significantly and meaningfully with those of human raters.

Compared to IEA and e-rater, PEG has the advantage of being conceptually simpler and less taxing on computer resources. PEG is also the better choice for evaluating writing style, as IEA returns grades that have literally nothing to do with writing style. IEA and e-rater, however, appear to be the superior choice for grading content, as PEG relies on writing quality to determine grades.

All three of these systems are proprietary and details of the exact process are not generally available. We do not know, for example, what variables are in any model nor their weights. 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. These criticisms are more problematic for PEG than for IEA and e-rater.

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. Because the systems are all based on normative data, the current state of the art does not appear conducive for scoring essays that call for creativity or personal experiences. The greatest chance of success for essay scoring appears to be for long essays that have been calibrated on large numbers of examinees and which have a clear scoring rubric.

Those who are interested in pursuing essay scoring may be interested in the Bayesian Essay Test Scoring sYstem (BETSY), being developed by the author based on the naive Bayes text classification literature (e.g., McCallum and Nigam, 1998). Free software is available for research use.

While recognizing the limitations, perhaps it is time for states and other programs to consider automated scoring services. We don't advocate abolishing human raters. Rather we can envision the use of any of these technologies as a validation tool with each essay scored by one human and by the computer. When the scores differ, the essay would be

flagged for a second read. This would be quicker and less expensive than current practice.

We would also like to see retired essay prompts used as instructional tools. The retired essays and grades can be used to calibrate a scoring system. The entire system could then be made available to teachers to help them work with students on writing and high-order skills. The system could also be coupled with a wide range of diagnostic information, such as the information currently available with IEA.

Key web sites

PEG - http://134.68.49.185/pegdemo/ref.asp
IEA - http://www.knowledge-technologies.com/
E-rater - http://www.ets.org/research/erater.html
BETSY - http://ericae.net/betsy/

References and Recommended Reading

Blok, H., & de Glopper, K. (1992). Large-scale writing assessment. In L. Verhoeven (Ed.), J. H. A. L. De Jong (Ed.), *the Construct of Language Proficiency: Applications of Psychological Models to Language Assessment,* pp. 101-111. Amsterdam, Netherlands: John Benjamins Publishing Company.

Breland, H. M., & Lytle, E. G. (1990). Computer-assisted writing skill assessment using WordMAP. ERIC Document Reproduction Service No. ED 317 586.

Burstein, J., K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder, and M.D. Harris (1998). Automated scoring using a hybrid feature identification technique. In the *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, August, 1998. Montreal, Canada. Available on-line: http://www.ets.org/reasearch/aclfinal.pdf

Burstein, J. (1999). Quoted in Ott, C. (May 25, 1999). Essay questions. *Salon*. Available online: http://www.salonmag.com/tech/feature/1999/05/25/computer_grading/

Chung, G. K. W. K., & O'Neil, H. F., Jr. (1997). Methodological Approaches to Online Scoring of Essays. ERIC Document Reproduction Service No. ED 418 101.

Fan, D. P., & Shaffer, C. L. (1990). Use of open-ended essays and computer content analysis to survey college students' knowledge of AIDS. *College Health, 38,* 221-229.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, *25*, (2&3), 285-307.

Jones, B. D. (1999). Computer-rated essays in the English composition classroom. *Journal of Educational Computing Research, 20(*2), 169-187.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Foltz, P. W, & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259-284.

McCallum, A. and K. Nigam (1998). A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on "Learning for Text Categorization". Available on-line:
http://citeseer.nj.nec.com/mccallum98comparison.html

McCurry, N., & McCurry, A. (1992). Writing assessment for the twenty-first century. *Computer Teacher, 19,* 35-37.

Page, E. B. (1966). Grading essays by computer: Progress report. Notes from the 1966 Invitational Conference on Testing Problems, 87-100.

Page, E.B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62(2), 127-42.

Page, E. B., Poggio, J. P., & Keith, T. Z. (1997). Computer analysis of student essays: Finding trait differences in the student profile. AERA/NCME Symposium on Grading Essays by Computer.

Rudner, L.M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research & Evaluation*, 3(3). Available online: http://pareonline.net/getvn.asp?v=3&n=3.

Vetterli, C. F., & Furedy, J. J. (1997). Correlates of intelligence in computer measured aspects of prose vocabulary: Word length, diversity, and rarity. *Personality and Individual Differences, 22(*6), 933-935.

Whissel, C. (1994). A computer program for the objective analysis of style and emotional connotations of prose: Hemingway, Galsworthy, and Faulkner compared. *Perceptual and Motor Skills, 79*, 815-824.

Whittington, D., & Hunt, H. (1999). Approaches to the computerized assessment of free text responses. *Proceedings of the Third Annual Computer Assisted Assessment Conference,* 207-219. Available online: http://cvu.strath.ac.uk/dave/publications/caa99.html.

Wresch, W. (1993) The Imminence of Grading Essays by Computer - 25 Years Later. *Computers and Composition,* 10(2), 45-58. Available online: http://corax.cwrl.utexas.edu/cac/archiveas/v10/10_2_html/10_2_5_Wresch.html.

Descriptors: Essays; Constructed Response; Scoring; Artifical Intelligence

Citation: Rudner, Lawrence & Phill Gagne (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation*, 7(26). Available online: http://PAREonline.net/getvn.asp?v=7&n=26.