DOCUMENT RESUME

ED 432 589 TM 029 965

AUTHOR Shermis, Mark D.; Koch, Chantal Mees; Page, Ellis B.; Keith, Timothy Z.; Harrington, Susanmarie

TITLE Trait Ratings for Automated Essay Grading.
PUB DATE 1999-04-21
NOTE 30p.; Some figures may not reproduce clearly.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Automation; College Students; Construct Validity; *Essays; *Grading; High School Students; High Schools; Higher Education; Interrater Reliability; Test Scoring Machines
IDENTIFIERS *Project Essay Grade

ABSTRACT

This study used Project Essay Grade (PEG) to evaluate essays both holistically and with the rating of traits (content, organization, style, mechanics, and creativity) for Web-based student essays that serve as placement tests at a large Midwestern university. In addition, the use of a TopicScore, or measure of topic content for each assignment, was incorporated into the PEG model to determine how well it would correlate with the five traits. In the first experiment, essays from 500 students were used to create statistical predictions for the PEG software. In the second experiment, the ratings from 300 essays were compared with ratings from 6 human judges. The interjudge correlation of the raters was only 0.51, but the prediction of all 6 judges, in the blind test, reached 0.83 for the PEG program. Of the five traits, "content" and "creativity" had the highest interjudge correlations. The new TopicScore correlated most highly with content, providing some measure of PEG's construct validity. The PEG software was an efficient means of grading the essays, with approximately six documents graded per second. (Contains 3 figures, 6 tables, and 15 references.) (Author/SLD)

Running Head: Automated Essay Grading

Trait Ratings for Automated Essay Grading

Mark D. Shermis

Chantal Mees Koch

Indiana University Purdue University Indianapolis

Ellis B. Page

Duke University

Timothy Z. Keith

Alfred University

Susanmarie Harrington

Department of English

Indiana University Purdue University Indianapolis

Abstract

This study employed Project Essay Grade (PEG) to evaluate essays both holistically and also with the rating of Traits (*Content, Organization, Style, Mechanics,* and *Creativity)* for web-based student essays that serve as placement tests at a large Midwestern university. In addition, the use of a TopicScore, or measure of topic content for each assignment, was incorporated into the PEG model to determine how well it would correlate with the five traits. The results of two combined experiments are reported, all based on random selection from about 800 essays. In the first experiment, the essays of 500 students were used to create statistical predictions for the PEG software. PEG used three major experimental strategies (some combining all observed variables) for these 500 essays. In the second experiment, the ratings from a separate, random sample of 300 essays were used to compare the ratings of six human judges against those generated by PEG. The inter-judge correlation of the human raters was only .51. But the prediction of all 6 judges, in the blind test, reached .83 for the PEG program. Of the five traits, *Content* ($\underline{r}$ = .54) and *Creativity* ($\underline{r}$ = .53) had the highest inter-judge correlations, even higher than those given for the overall *Holistic* ($\underline{r}$ = .51) rating. The new *TopicScore* measure correlated most highly with the trait of *Content* ($\underline{r}$ = .54), providing some evidence of PEG's construct validity. Finally, the PEG software was an efficient means for grading the essays with a capacity for approximately 6 documents graded every second. No delays in processing were observed by providing the additional trait scores. Other potential feedback measures for use in writing courses are discussed.

Trait Ratings for Automated Essay Grading

Introduction

Project Essay Grade (PEG) (Page, 1994) refers to computer software designed to evaluate written English text. The algorithms used to accomplish this are based on stable statistical models configured specifically for the type of writing to be assessed. For example, Shermis, Mzumara, Olson, & Harrington (1998) described one project involving the use of PEG technology for the evaluation of an English placement test, with impressive results. This experiment was noteworthy in demonstrating the applicability of PEG in a web-based testing environment with a turnaround processing time of seconds. A follow-up study by Mzumara and his colleagues (1998) showed that holistic PEG ratings had significantly better predictive validity than the ratings provided by human raters, using grades as the outcome variable. PEG has also been evaluated on nationally-normed tests that have significant writing components such as the GRE (Petersen & Page, 1997), Praxis (Page & Petersen, 1995), and NAEP (Page, Poggio, & Keith, 1997).

Because of recent publicity surrounding the use of automated essay graders (McCollum, 1999), it may be helpful to discriminate among the various available products. All graders use some sort of parser that partitions the writing into a taxonomic framework. For example, a parser might be programmed to identify the number of adverbs or unique words used in a sample of writing. Some parsers can detect the use of logical formulations when, for instance, a writer presents an argument using the syntax of "First,…second,…, and finally…".

Most of the automated text graders incorporate the evaluation of content as a significant component of their predictions by employing the use of keywords or their synonyms. For example, Landauer and his colleagues (1998) have applied "latent semantic analysis" (LSA) to their text grading engine as a way to determine the Euclidean distance between the desired and actual responses. LSA uses empirical ratings from judges as the basis for determining the distances among words. It also permits the grader to set up a desired answer by having it evaluate sections of text from a third source (e.g., a textbook) in setting the parameters for a desired outcome.

Text-graders that emphasize the evaluation of content have a number of important uses and will play a major role in the evaluation of aptitude and achievement tests. There are, however, three general criticisms that have been leveled at these grading engines. First, the claim that the computer can actually "understand" the text is not accurate. It is possible to write on a prompt using appropriate keywords and synonyms, but still lack a comprehensible answer. Consider the following:

> Queen America sailed to Santa Maria with 1492 ships. Her husband, King Columbus, looked to the Indian explorer, Nina Pinta, to find vast wealth on the beaches of Isabella, but would settle for spices from the continent of Ferdinand.

Of course the answer above is designed to be ridiculous, though some parsers might give it a high score for content since the passage contains many of the keywords associated with Columbus' discovery of North America. The counter-argument here is that if students are clever enough to creatively construct a response such as that listed above, they could probably generate a correct essay as well. The problems with automated grading most likely will stem from responses near the desired answers rather than those on the margins. Thus, some

researchers recommend a combination of human and machine graders, whenever grading has "high stakes" for the writer. (In such programs, it is customary to employ *two* human judges today.)

A second criticism leveled against graders that emphasize content has to do with the effort required to set up the models for each of the prompts. Most of the automated text graders use some sort of regression approach in setting up the statistical models. Depending on how many variables are involved, these models may require thousands of cases in order to derive stable regression weights. The implication here is that the methodology limits the grader's practical utility to large-scale testing operations where such data collection is feasible.

Finally, an over-arching concern with "content-heavy" automated text graders has to do with the effective use of one's assessment time in having individuals produce essays where *correct writing* is the most important attribute being evaluated. Most writing teachers emphasize the rhetorical aspects of the communication process such as the use of logic and persuasion in communicating one's views. In fact, some instructors purposefully assign essays that have no correct answer as a way to emphasize the building of writing skills (e.g., should students participate in some form of compulsory national service?). If the answer correctness is important, then other testing formats would probably be more efficient.

The PEG software distinguishes itself in that while it can evaluate content, most of the model development has been directed towards the creation of a general writing model, one that can effectively evaluate written work across a variety of prompts or topics. The software used here is not "intelligent" in the sense that it pretends to understand the content of the essay, but rather emulates the behavior of raters. "PEG is not aimed so much at AI [Artificial

Intelligence], … as at 'IA'--'Intelligent Assistance.' PEG won't replace the English teacher, but will serve as a useful, time-saving check on quality in writing" (Page, Lavoie, & Keith, 1996).

Keith (1998) evaluated the model generated for the IUPUI English placement exam and applied it to other data sets as part of an overall evaluation of the construct validity of PEG. The other samples included tests of PEG for the GRE (Petersen & Page, 1997), Praxis (Page & Petersen, 1995), and NAEP (Page et al., 1997), among others. He found that using the placement exam model to predict rater outcomes performed as well or better than the models originally developed for each study alone. His conclusion was that it would only be a matter of time before a general writing model would be developed for the assessment of other writing formats. Additional work is currently underway to determine whether or not a general writing model can be developed for the assessment of formats other than tests (e.g., electronic portfolio documents).

The initial applications of automated text graders will be to provide assistance in the summative evaluation of written work. However, the automated text grading has its greatest potential in providing students with formative feedback about areas of strength and weakness. Towards that end Page, Keith, and LaVoie (1996) identified five traits that typically emerge from the ratings of essays. These include: *content*, *organization*, *style*, *mechanics*, and *creativity*. Providing students with feedback on these dimensions has not only the potential to provide more detail in a summative evaluation, but could indicate to instructors areas where more writing development might be emphasized.

Page, Poggio, & Keith (1997) studied whether human raters have higher levels of agreement when ratings are provided holistically or when traits are explicitly identified. Eight judges were asked to provide both trait and holistic ratings on 495 essays in the NAEP essays from 1988. The results showed that the agreement coefficients for holistic ratings among human judges were higher than their corresponding trait agreement ratings. Moreover, for both holistic and trait ratings, PEG had coefficients that were as good or considerably higher than the ratings between two judges, or more.

The present study was designed as a larger scale replication of the Page, Poggio, & Keith effort (1997). In addition to a focus on the reliability of holistic versus trait ratings, however, some interest was devoted to assessing the additional variance explained by incorporating the content capabilities of PEG.

Hypotheses

1. Agreement on holistic ratings will be as high or higher than any individual trait rating.
2. External measures of topic adherence will be related to the "content" rating on the trait scale.
3. The addition of trait ratings will not add any important processing time to PEG evaluations.

Method

Participants

Study 1 (Forming the Model). Participants were 500 students drawn from a large Midwestern university and a suburban high school. All entering students at the university are required to take tests of math,

reading, and written English essays in order to be placed in appropriate courses. Students from the high school were participating as part of an experimental program to determine if taking placement tests at the secondary school produces a higher proportion of better prepared college students (Shermis, 1997; Shermis, Mzumara, Lillig, & Brown, 1997).

Study 2 (Test sample). Participants were 300 students drawn from the same large Midwestern university and suburban high school as in Study 1.

Instruments

English Placement Exam. The English placement exam is a one-hour exam that asks students to write an essay that explains and supports their opinion on a current social issue. Students have a choice of two questions, each providing a brief explanation of the issue for the context in which the test question is posed (Harrington, Shermis, & Rollins, 1998). Students are also asked to evaluate their answer and explain what changes they might make, had they the time to do so.

The scoring system uses a range extending from 1 (poor) to 6 (excellent). Raters used a web-based form to fill out their evaluations (see Figure 3), first providing the ratings on the traits followed by the holistic rating. The order of traits was presented in a fixed format. Raters were blind to the evaluations of others.

While placement rates may vary from year to year, on the whole 60% of the students taking the test are placed into first-year composition, 35% are placed into basic writing, and roughly 5% are placed into either honors, English as a Second Language (ESL), or other special courses. Most ratings are provided by faculty who teach first-year composition and basic writing; honors placements are made by faculty who teach honors courses. Based on earlier work with

holistically scored essays, the median correlation among six raters was r = .62 (Shermis et al., 1998).

Research on comparable scoring systems at other institutions suggests that training and shared teaching expertise creates acceptable levels of inter-judge agreement (cf. Smith, 1993; White, 1995). The predictive validity of the test has been computed with correlations in the low .20's (Mzumara et al., 1998) with course grades as an outcome variable.

Procedure

How PEG works. In much the same way as one might develop a statistical model with observed and latent variables, the evaluation of writing could be expressed in terms of *trins* and *proxes*. Trins are in*trins*ic variables of interest such as diction, fluency, and grammar (Page & Petersen, 1995). Proxes are from ap*prox*imations, that is, the observed variables with which the computer works, and are statistically calculated in the various writing samples. Examples of proxes might include the length of the essay or average word length. The statistical model for evaluating essays is formulated by optimizing the regression weights for the proxes and predicting rater averages of these trins. The rating generated by the statistical model is, in turn, compared against a new or test sample of average ratings among human judges.

Study 1. In our study just completed, students entered their essays using a screen (or web form) similar to that shown in Figure 1. Once the essay was completed, students submitted the text to a database that is controlled by a web server. Figure 2 illustrates a typical database entry. Six raters drawn from a pool of 15 instructional faculty provided their assessments on line by reading the essays and scoring

them. Essays from the first sample were analyzed to form the statistical model as part of Study 1. In this study, the proxes were identified and optimally weighted using the average judges' ratings as the outcome variable.

Insert Figures 1 & 2 About Here

The topic descriptions were scanned for vocabulary, and an algorithm stored the key words into an expanded new list. This list produces a new variable called a *TopicScore*. Related forms are added, so that PEG will recognize word-transformation from adjective to noun, or verbs changed in tense, etc.

Study 2. In the second study, essays were first sent to the database and rated by the instructors as before. PEG automatically queried the database to determine if new essays were present. If so, it transferred and processed the text, and returned the PEG score to the database. PEG scores are generated both as whole numbers (with a mean of 70 and standard deviation of 10) and z-scores.

Results

The statistical model sample (N = 500) consisted of 46.6% males and 53.0% females (.4% missing); 80.2% Whites and 17.6% Non-whites (2.2% missing). Since the assessment is a placement test, it was not surprising to see the high distribution in lower class levels: 87.6% freshmen, 4.6% sophomores, 1.2% juniors, and 6.6% other/missing. The average age was 22.65 with a standard deviation of 6.91. Table 1 shows the demographic characteristics of the sample by site (University or High School). The gender and ethnicity demographics closely match that of both participating institutions.

Table 2 shows the background characteristics of the test sample and is roughly similar to the characteristics specified in the statistical model sample. The test sample (N = 300) consisted of 46.3% males and 53.0% females (.7% missing); 77.7% Whites and 19.3% Non-whites (3.0% missing). The class distribution was 87.7% freshmen, 5.7% sophomores, 1.0% juniors, and 5.7% other/missing. The average age was 22.42 with a standard deviation of 7.20. Table 2 shows the demographic characteristics of the sample stratified by site (University or High School) and with demographic variables again being similar to both institutions.

Insert Tables 1 & 2 About Here

With respect to the efficacy of trait versus holistic ratings, Table 3 summarizes the median correlation among raters across all five traits and the overall rating for the test sample. Within the sample *content* (r = .54) had the highest agreement, followed by *creativity* (r = .53), *holistic* (r = .51), *mechanics* (r = .51), *organization* (r = .50), and style (r = .48). None of the differences in correlations are statistically significant. The next section of the table provides the PEG predictions of each trait and the holistic ratings, based on the Spearman-Brown Formula. For example, the median correlation among the raters for the overall holistic ratings was r = .51. The correlation between PEG's ratings and the average ratings among the six raters was r = .83. PEG had statistically significant improvements in predictive power across all five traits and the overall holistic score. The final column shows the power of PEG in comparison with three judges or more—even with four or more judges.

Insert Table 3 About Here

Correlations between holistic ratings provided by PEG and three pairings of judges are represented in Table 4. As can be seen in this table, the holistic ratings predicted by PEG tend to be more highly correlated with the three different judge paired ratings than the judge pairs are intercorrelated. In the sample, PEG predicted ratings are correlated (r=.75) with the judge pairs and (r=.72) with each other, although this difference is not statistically significant.

Insert Table 4 About Here

A confirmatory factor analysis (N = 300) was performed between the PEG ratings and three judge combinations described above. In this analysis, the "essay true score" represents the underlying latent trait of writing ability. As Figure 3 shows, PEG performed as well or better than the highest of the judge combinations.

Insert Figure 3 About Here

The correlations among raters were broken down by topic for the holistic rating. Essay topics rotated approximately every two weeks and 16 different topics were included in the model. Table 5 summarizes the median correlations by topic. The median (weighted) correlation by topic was r = .58 with a range between r = .40 to r = .72.

Insert Table 5 About Here

With regard to the relationship between the *TopicScores* and ratings across traits, the correlations across 500 essays, 6 judges, 19 topics, and 6 categories were calculated. These are presented in Table 6. As can be seen these assigned *TopicScores* correlated most highly with *Content,* followed by the overall, or *Holistic* rating.

Finally, the additional processing time required by PEG to generate the trait ratings was found to be negligible. On a Pentium II 250 MHz computer, PEG can process six essays per second. On a Pentium II 400 MHz machine, it can again process six essays per second. These are comparable to previous evaluations of CPU processing where only the holistic score was generated.

Insert Table 6 About Here

Discussion

In grading any papers, one of the most useful strategies is to notify the student where an essay is strong or weak. But no teachers exist who grade such traits in any uniform way. One of the greatest contributions of PEG is that it provides a reasonable and economical method for doing this.

In only one other modern data set has this been done: PEG used eight ratings of a full set of Traits, to grade nearly 500 essays from

the National Assessment of Education Progress (NAEP), for one grade {different grades} in U.S. secondary schools (Page et al., 1997). All of these writing exercises (within a NAEP year) were on a single topic. For that study, PEG employed eight qualified raters to grade these essays: on *holistic*, then on *content*, *organization*, *style*, *mechanics*, and *creativity*. Subsequent analyses showed that, not only were the NAEP essays powerfully graded overall, but the residuals from the traits (after subtracting the influence of *holistic*) showed they could be powerful discriminators within a student's writing style.

As a way to develop a more stable trait model, the present study replicated many of the features of Page's (1997) earlier work and the Shermis et al. (1998) study using the holistic ratings only. First, PEG once again performed statistically significantly better than human raters with an $\underline{r}$ =.51 between raters (for six raters) and an $\underline{r}$ = .83 for the average between the raters and PEG on the holistic ratings. Similar yields of improvement were made from the trait ratings as well. These results were a bit more dramatic than the differences observed in the Shermis (1998) study where the inter-judge correlation of the human raters was $\underline{r}$ = .62 and was $\underline{r}$ = .71 for the computer.

Based on the previous Page study (1997), the expectation was that the holistic ratings would have significantly higher intercorrelations than any one of the individual traits. Surprisingly, this was not the case. The highest correlations in the sample were found with ratings on *content* and *creativity*, followed by the *holistic* ratings. It turned out that the ratings for the more mechanical aspects of rating (as defined by the traits) had turned out to have the lowest inter-judge correlations. None of these differences among the traits or holistic ratings were particularly large nor were they significantly different

from one another. What made the *content* and *creativity* trends particularly noteworthy was that, in contrast to the earlier NAEP study, the essays included over 16 different topics. Moreover, none of the essays included in the analysis had a "correct" response. Most of them were rhetorical exercises in which writers had to take a stand, logically defend it, and then speculate how they could have provided a more persuasive argument (e.g., should gambling be made legal in this state?). One could speculate that the numbers for *content* would have been even higher had the number of topics been reduced and/or there had been a correct response.

The lower correlations in the sample associated with *style*, *mechanics*, and *organization* were a bit unexpected. Many writing instructors refer to these traits as the "superficial" aspects of composition, yet there appeared to be less agreement on these components than on the "deeper" traits of *content* and *creativity*. If in fact these traits are valued less, then the relative lack of agreement may simply be a reflection of rater inattention to these aspects of writing. Again, the differences among the traits were not statistically significant, so it could be that these traits were all equally salient from the raters' viewpoints. We plan on conducting a more in depth analysis of rater disagreements at a later point.

In spite of the higher than expected agreement coefficients on *content*, PEG's *TopicScore* correlated more highly with this trait than did any of the other available ratings, including the *holistic* rating. This result, along with Keith's earlier work on the construct validity of PEG (1998), suggests that PEG is correctly targeting the underlying *intrinsic* structures of writing valued most by raters.

The addition of the PEG trait scores did not add any perceptible delay in the server's processing of the essays. The additional

information does not come at any higher processing cost than does the holistic statistical model.

Is the investment of time worth it? Most writing teachers tell us that what they'd like to provide students more comprehensive feedback, but can't because their time is too constrained. Ideally, they'd like to give students a narrative that picks out a sample of what they've done well, a sample of what wasn't particularly impressive, a numerical summary of the student's strengths and weaknesses, and an overall grade. But this dream is a few years off. In the meantime, instructors (or test administrators) can provide a summary of trait ratings predictable from this research.

References

Harrington, S., Shermis, M. D., & Rollins, A. (1998). The influence of word processing on English placement test results . Indianapolis, IN: Indiana University Purdue University Indianapolis.

Keith, T. Z. (1998). Construct validity of PEG. Paper presented at the American Educational Research Association, San Diego, CA.

Landauer, T., Laham, D., & Foltz, P. (1998). The Goldilocks principle for vocabulary acquisition and learning: Latent Semantic Analysis theory and applications. Paper presented at the American Educational Research Association, San Diego, CA.

McCollum, K. (1999, January 29, 1999). Computers will help grade essays on Graduate Management Admission Test. The Chronicle of Higher Education, 44, A30.

Mzumara, H. R., Shermis, M. D., & Fogel, M. (1998). Validity of the IUPUI placement test scores for course placement: 1997 - 1998 . Indianapolis: IUPUI Testing Center.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. Journal of Experimental Education, 62(2), 27-142.

Page, E. B., Lavoie, M. J., & Keith, T. Z. (1996, April). Computer grading of essay traits in student writing. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. Phi Delta Kappan, 76(7), 561-565.

Page, E. B., Poggio, J. P., & Keith, T. Z. (1997, March). Computer analysis of student essays: Finding trait differences in the

student profile. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Petersen, N. S., & Page, E. B. (1997). New developments in Project Essay Grade: Second ETS blind test with GRE essays. Paper presented at the American Educational Research Association, Chicago, IL.

Shermis, M. D. (1997, March). Recent developments in college placement testing: Assessments via the World Wide Web. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Shermis, M. D., Mzumara, H. R., Lillig, C., & Brown, M. (1997, August). Computerized adaptive testing through the World Wide Web. Paper presented at the annual meeting of the American Psychological Association, Chicago, IL.

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington, S. (1998). On-line grading of student essays: PEG goes on the web at IUPUI. Paper presented at the American Educational Research Association, San Diego, CA.

Smith, W. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. M. Williamson & B. Huot (Eds.), Validating holistic scoring for writing assessment: theoretical and empirical foundations (pp. 142-205). Cresskill, NJ: Hampton Press.

White, E. (1995). Teaching and assessing writing. (2nd ed.). San Francisco: Jossey-Bass.

Author Notes

Correspondence concerning this article should be addressed to Mark D. Shermis, IUPUI Testing Center, 620 Union Drive, Indianapolis, IN 46202-5168. Electronic mail may be sent via Internet to MShermis@IUPUI.Edu. This work would not have been possible without the generous contributions of Clif Marsiglio of the IUPUI Testing Center and Vicki Hale of the IUPUI English Department.

Table 1.

Demographic Characteristics of Statistical Model Sample (N = 500).

Variable	Location			
	University N = 494		High School N = 6	
	98.8%		1.2%	
Gender				
Male	46.6		50.0	
Female	53.0		50.0	
Missing	0.4			
Ethnicity				
White	80.8		33.3	
Non-White	17.4		33.3	
Missing	1.8		33.3	
Class Level				
Freshman	88.1		NA	
Sophomore	4.7		NA	
Junior	1.2		NA	
Senior	0.0		NA	
Other/Missing	3.0		NA	
	Mean	SD	Mean	SD
Age	22.7	6.9	NA	NA

NA = Not Ascertained

Table 2.

Demographic Characteristics of the Sample Which Formed the Statistical Test (N = 300).

Variable	Location			
	University N = 296		High School N = 4	
	98.7		1.3%	
Gender				
Male	46.4		50.0	
Female	52.9		50.0	
Missing	0.7			
Ethnicity				
White	78.3		50.0	
Non-White	19.3		25.0	
Missing	2.4		25.0	
Class Level				
Freshman	89.2		NA	
Sophomore	4.7		NA	
Junior	1.0		NA	
Senior	0.0		NA	
Other/Missing	5.1		NA	
	Mean	SD	Mean	SD
Age	22.4	7.2	NA	NA

NA = Not Ascertained

Table 3.

Mean Correlation Between PEG and Judges across the five traits along with the PEG predictions (N = 300).

Dimension	Mean Correlation Between Judges	PEG Prediction of Six Judges[1]	Est. Number of Human Judges
Holistic	0.512	0.830	3++
Content	0.546	0.844	4++
Organization	0.499	0.767	3+
Style	0.476	0.808	4+
Mechanics	0.506	0.778	3
Creativity	0.525	0.833	4+

[1]Based on the Spearman-Brown Formula

Table 4.

Correlation of Holistic Ratings with Judge Pair Combinations

	1	2	3	4
	Essays (N = 300)			
1. PEG	-	0.778	0.759	0.723
Prediction	-			
2. Judges 1 & 2		- -	0.712	0.759
3. Judges 3 & 4			- -	0.680
4. Judges 5 & 6				- -

PEG Agreement with Pairs	Judge Pair Agreement
0.753	0.717

Table 5.

Median Correlations of Holistic Scores between 6 Human Raters by Topic Number (N = 300)

Topic	# of essays	Median r
Missing	8	--
51	130	.55
55	74	.56
56	18	.69
62	62	.57
64	93	.63
66	49	.40
67	23	.62
69	60	.59
72	6	.71
73	29	.53
75	35	.72
76	49	.59
77	18	.65
79	91	.51
80	19	.63
83	36	.69

Table 6.

Correlation of TopicScore with judge ratings of traits.

Dimension	r
Holistic	.516
Content	.538**
Organization	.474
Style	.481
Mechanics	.473
Creativity	.484

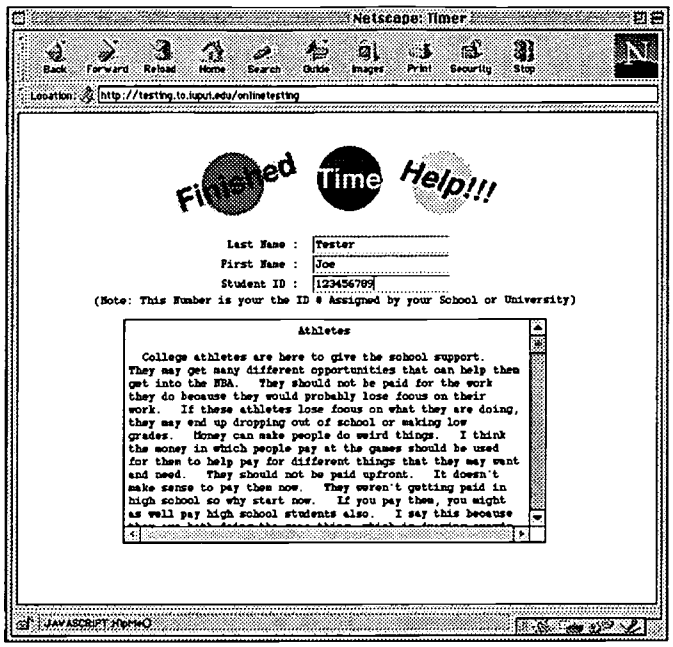
** $p < .01$

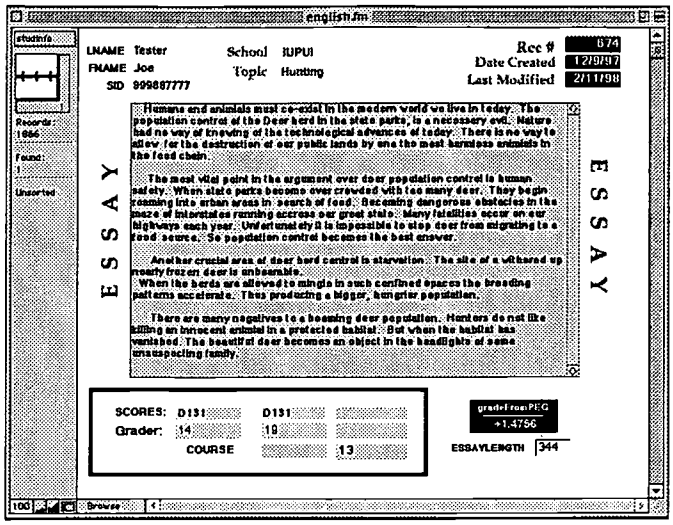
Figure Captions

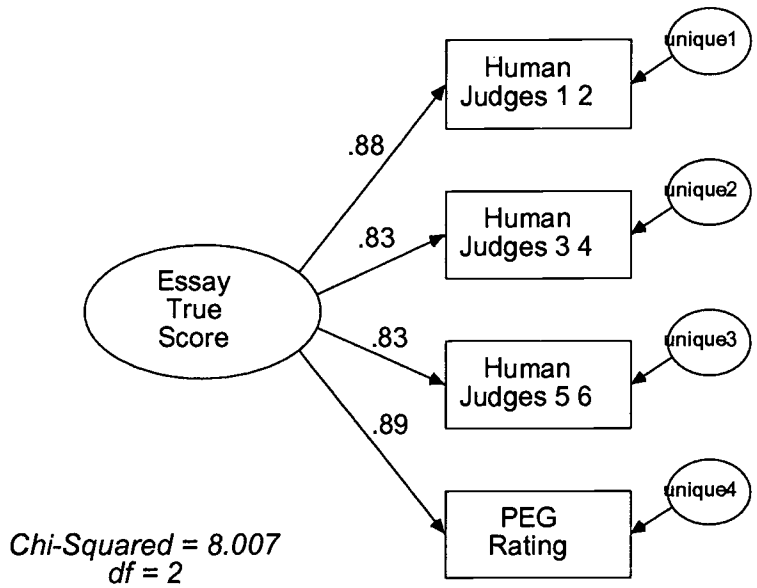
Figure 1. The web form used for the English written examination.

Figure 2. The database that stores the English written examination.

Figure 3. The results of a confirmatory factor analysis comparing computer ratings to pairs of human judges.




english.fm
studinfo
Records:
1886
Found:
1
Unsorted
LNAME Tester
FNAME Joe
SID 999887777
School IUPUI
Topic Hunting
Rec # 674
Date Created 12/9/97
Last Modified 2/11/98
ESSAY
Humans and animals must co-exist in the modern world we live in today. The population control of the Deer herd in the state parks, is a necessary evil. Nature had no way of knowing of the technological advances of today. There is no way to allow for the destruction of our public lands by one the most harmless animals in the food chain.
The most vital point in the argument over deer population control is human safety. When state parks become over crowded with too many deer. They begin roaming into urban areas in search of food. Becoming dangerous obstacles in the maze of interstates running across our great state. Many fatalities occur on our highways each year. Unfortunately it is impossible to stop deer from migrating to a food source. So population control becomes the best answer.
Another crucial area of deer herd control is starvation. The site of a withered up mostly frozen deer is unbearable.
When the herds are allowed to mingle in such confined spaces the breeding patterns accelerate. Thus producing a bigger, hungrier population.
There are many negatives to a booming deer population. Hunters do not like killing an innocent animal in a protected habitat. But when the habitat has vanished. The beautiful deer becomes an object in the headlights of some unsuspecting family.
ESSAY
SCORES: D131 D131
Grader: 14 19
COURSE 13
gradefromPEG
=1.4756
ESSAYLENGTH 344
100
Browse



Chi-Squared = 8.007
df = 2
p = .018
GFI = .986
TLI = .979
CFI = .993

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM029965

REPRODUCTION RELEASE

(Specific Document)

AERA

I. DOCUMENT IDENTIFICATION:

Title: Trait Ratings for Automated Essay Grading

Author(s): Mark D. Shermis, Chantal M. Koch, Ellis B. Page, Timothy Z. Keith, Susanmarie Harrington

Corporate Source: IUPUI Testing Center

Publication Date: 4/21/99

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

[X]

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

[ ]

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

[ ]

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Mark D. Shermis	Printed Name/Position/Title: Director
Organization/Address: IUPUI Testing Center, 620 Union Dr., Ste. 6003, Indianapolis, IN 46202-5167	Telephone: 317-278-2286 FAX: 317-274-3400
	E-Mail Address: MShermis@IUPUI.Edu Date: 6/9/99

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com