A Comparison of the Accuracy of Automated Essay Scoring Using Prompt-Specific and Prompt-Independent Training

Cathy Mikulas
Kevin Kern
Vantage Learning

110 Terry Drive Suite 100
Newtown, PA 18940
www.vantagelearning.com

Paper to be presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA – April 2006

Correspondence regarding this presentation may be directed to Cathy Mikulas at cmikulas@vantage.com.

A Comparison of the Accuracy of Automated Essay Scoring Using Prompt-specific and Prompt-independent Training

Background

IntelliMetric™ is an automated essay scoring tool developed by Vantage Learning that uses Artificial Intelligence, Natural Language Processing, and Statistics in its scoring of essays. The development of IntelliMetric™ began in the 1980s. Since 1998, it has been used successfully to score open-ended essay-type assessments. IntelliMetric™ was the first commercially successful tool able to administer open-ended questions and provide feedback to students in a matter of seconds.

Hundreds of studies have been conducted to evaluate the quality of IntelliMetric™ scoring (see Shermis & Burstein, 2002). Agreement rates (exact, adjacent, and discrepant) with expert human scorers and correlations between IntelliMetric™ and human scores are the most common methods of evaluating the quality of IntelliMetric™ and other automated essay scoring engines. In essence, the expert human scoring is a baseline for the quality of automated essay scoring engines. IntelliMetric™ has been shown to be as accurate as or more accurate than expert scorers. In other words, IntelliMetric™ is able to agree with expert human scorers more often than experts agree with each other.

An Overview of IntelliMetric™ Training and Validation

How Does IntelliMetric™ Score Essays?

IntelliMetric™ is an intelligent scoring system that emulates the process carried out by human scorers. IntelliMetric™ is theoretically grounded in a cognitive model often referred to as a "brain-based" or "mind-based" model of information processing and understanding. IntelliMetric™ draws upon the traditions of Cognitive Processing, Artificial Intelligence, Natural Language Understanding and Computational Linguistics in the process of evaluating written text.

The system must be "trained" with a set of previously scored responses with known scores as determined by experts. These papers are used as a basis for the system to infer the rubric and judgments of the human scorers. The systemic interaction of over 400 semantic, syntactic and discourse level features of text is examined by IntelliMetric™ and categorized with each rubric score point. The IntelliMetric™ system classifies the characteristics of the responses associated with each score point and applies this intelligence to score essays with unknown scores.

Key Principles. IntelliMetric™ is based on a brain-based model of understanding and follows five key principles. They are:

1. **IntelliMetric™ is modeled on the human brain.** A neurosynthetic approach is used to reproduce the mental processes used by human experts to score and evaluate written text.

   Many mark the formal beginning of inquiry into how the mind creates meaning with William James' (1890) fundamental work in association. Inquiry into understanding continued through the early part of the twentieth century with the behavioral movement. Research then moved towards a more cognitive understanding of meaning with the early work of Joos (1950) in language understanding and Osgood, Suci, and Tannenbaum's (1957) landmark work *The Measurement of Meaning.* Understanding how we understand has been the holy grail of cognitive science. Minsky (1986) captured the perspective embodied by IntelliMetric™ in his view of the brain presented in *The Society of Mind*; here, understanding is seen as the result of billions of interacting subprograms, each doing simple computations. The cognitive scientific approach to understanding continued to grow throughout the latter part of the twentieth century. Most recently Baum's (2004) work has extended this search and has produced an integrated view of meaning.

2. **IntelliMetric™ is a learning engine.** IntelliMetric™ acquires the information it needs by learning how to evaluate writing based on examples that have already been scored by experts. IntelliMetric™ is able to handle inconsistencies within the training set and develop its own scoring. This can be seen directly through the IntelliMetric™ rescoring of the training set that was provided to teach IntelliMetric™ how to score essays for a particular prompt. IntelliMetric™ will not assign the exact same scores as provided in the training set because its scoring model is a reflection of more than the individual essays and scores.

3. **IntelliMetric™ is systemic.** IntelliMetric™ is based on a complex system of information that together yields a result that is much more than its component parts. Judgments are based on the overall pattern of information and the preponderance of evidence.

4. **IntelliMetric™ is inductive.** IntelliMetric™ makes judgments inductively rather than deductively. Judgments are made based on inferences built from "the bottom up" rather than "hard and fast" rules. In other words, IntelliMetric™ is not rule-based. It is not handed the rubric and rules about weighting certain features or domains. Rather, it is provided with a training set of hundreds of essays that were scored by experts. IntelliMetric™ then establishes its own system for scoring that enables it to predict human scores on new essays.

5. **IntelliMetric™ uses multiple judgments based on multiple mathematical models.** IntelliMetric™ is based on several different types of judgments using many types of information organized using sophisticated mathematical tools. Rather than using just one solution for automated essay scoring, IntelliMetric™ incorporates multiple methods of evaluation. These methods are referred to as "judges." Each judge predicts a score and those scores are optimized to yield the final IntelliMetric™ score. This is similar to the

human scoring process in which multiple scorers are used to yield the most accurate score for each essay.

How is IntelliMetric™ Trained to Score Essays?

IntelliMetric™ is trained in a similar manner to traditional human scorer training. In human scoring, the scorers are given detailed instruction on the rubric and its interpretation. Scorers are provided with a sampling of previously scored essays (often referred to as "anchors") accompanied with explanations of why each essay was given that particular score. The scorers are then able to score some essays on their own. After a few rounds of feedback and calibration, if the scorer is able to score new essays at a predetermined level of agreement with other scorers, the scorer is given an operational scoring assignment.

IntelliMetric™ is trained in much the same way as described above. IntelliMetric™ is given a set of approximately 300 anchor papers (the training set) as the basis for training. IntelliMetric™ learns the characteristics of the score scale through exposure to the training set, which has been scored by experts. In essence, IntelliMetric™ internalizes the pooled wisdom of scorers included in the training set.

Much like human scorers who are typically trained on each specific question or prompt, IntelliMetric™ modeling is also unique for each prompt. This process leads to high levels of agreement between the scores assigned by IntelliMetric™ and those assigned by human scorers.

Training Set Composition

Since IntelliMetric™ is an inductive system that categorizes characteristics of essays and associated scores from a training set, it is critical that the training set reflects the range and composition of the work that will be submitted under operational conditions. This is similar to the notion of field testing new multiple choice test questions; it is important to ensure that the field testers are representative of those who will be taking the test questions operationally. Without this match, the scores will not be valid. With essay scoring, the same is true.

Studies regarding IntelliMetric™ scoring have yielded the following inferences regarding the optimal composition of training sets:

- ***Include at least 300 training papers.*** Although accurate models have been constructed with as few as 50 training papers, an ideal training set consists of 300 or more papers.
- ***Provide sufficient coverage across each score point including the tails.*** For example, on a one to six scale it is important to include at least 20 papers defining the "1" point and the "6" point. The reason for this is the inductive nature of the modeling; without examples of a particular score point, the rubric is truncated.
- ***Include multiple raters if possible.*** Two or more scorers typically yield better results than one scorer. Any one scorer is subject to inconsistencies that will raise confusion during the model creation process.
- ***Use a six-point or larger scale.*** The variability offered by six as opposed to three- or four-point scales appears to improve IntelliMetric™ performance.

- ***Ensure the human scorers are well calibrated.*** While IntelliMetric™ is very good at eliminating "noise" in the data, ultimately the engine depends on receiving accurate training information. The adage "garbage in, garbage out" holds true with IntelliMetric™ modeling.

Under these conditions, IntelliMetric™ will typically outperform human scorers.

How Do We Know that IntelliMetric™ works?

The last step in the model creation process is to validate the scores from IntelliMetric™ from the new model. A set of validation essays that has been scored by expert human raters is withheld from the training process. Upon the creation of the model, the model is used to score this validation set of essays.

The means of the humans compared to the IntelliMetric™ model are compared, the agreement rates are calculated, and the Pearson correlation between the scores is calculated. If the means are not significantly different, the agreement rates meet a benchmark of acceptance (typically at least 70% exact agreement and less than 1% discrepant), and the Pearson correlations are .8 or higher, the model is typically approved for use. If the data does not support the use of the model, the scoring produced by the humans and IntelliMetric™ is carefully reviewed to determine the source of the discrepancies. When there are disagreements, either the humans are wrong, the computer is wrong, or they are both wrong. Based on the findings, appropriate adjustments are made and a new IntelliMetric model is created.

Literally thousands of IntelliMetric models have been developed and validated. The results show that IntelliMetric is able to reliably score essays submitted to the prompt at accuracy levels equal to or greater than expert humans.

While this prompt-specific IntelliMetric™ modeling has been shown to be very effective, the collection of training sets is often a burden for schools and districts. Because of this, there is a need for a more general model that could accurately score writing samples regardless of prompt. This study investigates the accuracy of a prompt-independent (content neutral) model as compared to a prompt-specific IntelliMetric™ model.

The Investigation

Traditionally, the creation of automated essay scoring (AES) models follows a prompt-specific training approach similar to that of expert scoring in which the scorers are given anchor papers and are calibrated to the rubric uniquely for each prompt. The AES training involves the collection of a substantial sample of previously scored essays for the prompt. The scoring technologies available support a more prompt-independent training in which the model is not concerned about the topic and is focused on the writing quality. These prompt-independent models remove the need for the collection of essays and also allow writing instruction administrators the flexibility to use a wide variety of prompts that have not yet been trained for scoring.

This study evaluates the scoring accuracy of models created under the conditions of prompt-specific and prompt-independent training in order to investigate the efficacy of these two approaches. While the technology for prompt-independent modeling is available and corresponding claims that this approach is as good as or better than prompt-specific modeling, it is critical that the stakeholders are aware of any differences in quality and subsequently any threats to the reliability and validity of the scores.

Data Source and Procedures

Three approaches to evaluating student writing were explored within this study. All three measurement methods evaluated student writing using a common 6 point rubric.

Human Expert Scoring. The criterion measure for comparing the prompt-specific and prompt-independent scoring approaches was the writing scores assigned by human expert scorers. The human expert score was the average of two expert raters. Any discrepancies were reviewed by a chief rater and the final human score in those cases is the score assigned by the chief reader.

Prompt-Specific AES Scoring. Approximately 300 middle-school students were administered a single persuasive writing prompt during the 2004-2005 academic year as part of the implementation and study of an instructional writing software program incorporating automated essay scoring. Human experts initially scored all of the essays submitted by the students. Then an automated essay scoring system (IntelliMetric™) was trained to score the essays. Approximately 250 of the sampled essays were used to train the sample and a randomly sampled set of 50 essays were withheld for validation.

Prompt-Independent AES Scoring. Approximately 3000 responses drawn from 10 middle school prompts including the prompt used as the prompt-specific model in this study were used as a basis for training the prompt-independent scoring model. The data set did not include the 250 responses used to train the prompt-specific model. The approximately 3000 responses along with the corresponding human expert scores were used as a basis for the IntelliMetric™ engine to "learn" to score responses. A set of 50 essays were withheld from training for validation.

Three studies were conducted using scores from these three approaches. The first study is an internal validation of the automated scoring model created to score the essays. This evaluation compares the IntelliMetric™ scores with those of the expert human scores. The second study is a cross-validation in which the first study was repeated 10 times to ensure consistency. The third study looks at the scoring accuracy of the prompt-specific and prompt-independent IntelliMetric™ models as compared to human scores for a set of 50 essays.

Analyses

Study 1: An Analysis of Prompt-Specific and Prompt-Independent Model Scoring Accuracy

When creating automated scoring models for research or operational use, the first step of the evaluation is to ensure that the model is functioning at an acceptable level internally. This validation of the data is determined through a comparison of human and computer scores of a subset of papers withheld from training. If the computer can score the validation set meeting rigorous agreement rate requirements, the model is acceptable for use. This first study provides the validation analysis of the two models used in the study.

Descriptive statistics. The means and standard deviation of the scores for the human scoring and the automated model were determined. These measures provide a useful comparison of the overall average score for humans and IntelliMetric™ and the amount of dispersion in each of the two data sets.

Agreement Analysis. The frequency with which IntelliMetric™ was in agreement with scores assigned by expert graders was calculated. The frequencies with which IntelliMetric™ and expert scorers agreed and did not agree were calculated.

Correlation Analysis. The Pearson r correlation between IntelliMetric™ classifications and scores assigned by expert graders was computed as a measure of the overall relationship between the two sets of data.

Results

Descriptive Statistics. The mean scores for the human ratings and the computer ratings for both scenarios were calculated. There were no differences found between the human and computer mean for either model ($p > 0.05$).

Table 1: Descriptive Statistics

	Prompt-Specific Model Validation Set		Prompt-Independent Model Validation Set	
	Mean	**SD**	**Mean**	**SD**
Human Expert	3.5	1.33	3.36	1.47
Computer	3.46	1.35	3.32	1.29

Agreement Rates. The extent to which the automated essay scoring models agreed with human scorers was examined. The prompt-specific IntelliMetric™ model yielded an exact agreement rate of 84% and an adjacent agreement rate of 100% while the prompt-independent model yielded rates of 62% and 94% respectively.

Correlation Analysis. The correlation between the automated essay scoring model scores and human expert scores was examined. The Pearson r correlations for the two models were .96 and .86 respectively. The two correlations were significantly different based on a z-test ($z=3.17$; $p<.01$).

Study 2: Cross Validation

In order to verify the stability/accuracy of the Study One results, the study was repeated using a 10 subsample cross-validation approach. The identical data source, measurement and study procedures were employed.

Procedures

For both the prompt-specific and prompt-independent AES scoring models, 10 random samples of 50 responses were selected. For each of the 10 sets of 50 responses, the human scores were compared to the AES model results. The agreement rates and correlations for the two scoring models were calculated for 10 sets of 50 essays.

Results

The prompt-specific results revealed Pearson r correlations ranging from .91 to .96; the prompt-independent results showed Pearson r correlations ranging from .86 to .89. As an index for comparison, the average correlation across the 10 sub sample validations was compared. The average correlation across the Prompt-specific models was .93, while the average across the prompt-independent sets was .88.

Table 2: Prompt-Independent Summaries Cross Validation (N=50)

Response Set	Exact	Adjacent	Discrepant	Pearson r
1	54%	46%	0%	0.89
2	62%	36%	2%	0.89
3	52%	48%	0%	0.89
4	54%	42%	4%	0.88
5	60%	36%	4%	0.87
6	62%	36%	2%	0.89
7	50%	50%	0%	0.88
8	48%	50%	2%	0.86
9	56%	44%	0%	0.89
10	58%	40%	2%	0.89

Table 3: Prompt-Specific Summaries Cross Validation (N=50)

Response Set	Exact	Adjacent	Discrepant	Pearson r
1	82%	18%	0%	0.95
2	74%	26%	0%	0.92
3	78%	22%	0%	0.94
4	84%	16%	0%	0.96
5	70%	30%	0%	0.91
6	74%	26%	0%	0.93
7	76%	24%	0%	0.93
8	74%	24%	2%	0.92
9	84%	16%	0%	0.95
10	74%	26%	0%	0.93

Study 3: A Comparison of Prompt-Specific and Prompt-Independent AES Scores to Expert Human Scores

The first two studies investigated the scoring accuracy of the two scoring models independent of each other. This third study investigated the scoring accuracy of each model when scoring the same set of student essays that were on topic for the prompt-specific model.

Procedures

A random set of 50 essays submitted to the middle school prompt that is the topic for the prompt-specific model was collected. Each essay was scored by an expert human scorer, the prompt-specific model, and the prompt-independent model. The results from the three scoring scenarios were compared.

Results

Descriptive Statistics. The mean scores for the human ratings and the computer ratings were calculated. There were no differences found between the human and computer mean for either model ($p > 0.05$).

Table 4: Descriptive Statistics

	Mean	SD
Human Expert	3.5	1.5
Prompt-Independent	3.44	1.4
Prompt-Specific	3.54	1.5

Agreement Rates. The extent to which the automated essay scoring models agreed with human scorers was examined. The prompt-specific IntelliMetric™ model yielded an exact agreement rate of 84% and an adjacent agreement rate of 100% while the prompt-independent model yielded rates of 74% and 100% respectively.

Correlation Analysis. The correlation between the automated essay scoring model scores and human expert scores was examined. The Pearson r correlations for the two models were .97 and .94 respectively. The two correlations were not significantly different based on a z-test ($z=1.72$; $p>.05$).

Discussion

The results confirm earlier findings that automated essay scoring (IntelliMetric™ in this study) can reliably score student written responses. However, the results also suggest that a prompt-specific training approach is superior to a more generic, prompt- independent approach. It appears that it is better to train to a specific prompt than to apply a more generic cross-prompt training approach. This is not surprising in light of what we know about human expert scoring. Writing experts generally advocate training human scorers to evaluate a specific prompt, rather than relying solely on generalized training. This stems largely from the recognition that a writer writes about something specific, for a specific audience and for a specific purpose. The prompt-specific training provides the IntelliMetric™ model with valuable information regarding what it means to be an essay that is legitimate.

While the prompt-independent approach produced a high degree of accuracy as determined through agreement with human scoring, the data needs to be interpreted with caution. The prompt-independent model will score the 50 essays from the third study consistently regardless of the prompt that is presented to the student. So while in this case the essay scores are accurate, the essay scores would be invalid for other prompts. This is a limitation to the prompt-independent model. While the prompt-independent approach may be able to score accurately with respect to human scoring, it is not able to detect whether or not the essay is on topic. This is a result of the notion of prompt-independent (and therefore content neutral) scoring.

Although the prompt-specific model is able to more accurately score essays submitted and determine whether the response is on topic, the prompt-independent method certainly provides some value for writing instruction. If there is no need to be able to detect students who are attempting to trick the system with an essay written to another topic and the score does not need to be as precise, the content-neutral model suits the situation. For assessment purposes in which decisions need to be made regarding placement, candidacy, or other important judgments, the data supports the use of the prompt-specific model.

The prompt-independent scoring model offered a reasonable approximation of the human expert score. This approach may be warranted within a range of instructional situations. The prompt-specific approach should be used for assessments and for instructional situations in which it is necessary to determine whether an essay is on topic or on task.

References

Baum, E.B. 2004. *What is Thought?* Cambridge, Massachusetts: MIT Press.

Blalock Jr, Hubert. 1972. *Social Statistics*. McGraw Hill. p. 405-407

Joos, M. 1950. Description of Language Design. *Journal of the Acoustic Society of America* 22:701-08.

Minsky, M. 1986. *Society of Mind.* Cambridge, Massachusetts: MIT Press.

Osgood, C.E., J. Suci, and P.H. Tannenbaum. 1957. *The Measurement of Meaning.* Urbana, Illinois: University of Illinois Press.

Schank, R.C. 1999. *Dynamic Memory Revisited.* Cambridge, England: Cambridge University Press.

Shermis, M. & Burstein, J. 2002. *Automated Essay Scoring. A cross-disciplinary perspective.* New Jersey: Lawrence Erlbaum Associates.

