The Effects of Inclusion of Native Speakers' Writing Samples on the Domain Scoring Accuracy of Automated Essay Scoring of Writing Submitted by Taiwanese English Language Learners

Paul Edelblut
Cathy Mikulas

Vantage Learning
110 Terry Drive
Newtown, PA 18940

www.vantagelearning.com
215-579-8390

Paper to be presented at
The 32nd Annual Conference of the International Association for Educational Assessment
Singapore, 21-26 May 2006

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

The Effects of Inclusion of Native Speakers' Writing Samples on the Domain Scoring Accuracy of Automated Essay Scoring of Writing Submitted by Taiwanese English Language Learners

Abstract

While the scoring accuracy of automated scoring of essays written in English has been established, more research is needed with regards to domain scoring for English Language Learners. This paper presents findings regarding the effects of training set composition on the domain (Focus and Meaning, Content and Development, Organization, Language Use and Style, and Mechanics and Conventions) scoring accuracy of essays submitted by Taiwanese students scored by an automated essay scoring system. Typically, each scoring model created is based on a set of previously scored essays. This study compares the accuracy of scoring the same set of essays written by Taiwanese students using two different models: one model using blended native and ELL essays and one using a set of entirely ELL essays. While both models yielded adjacent agreement rates from 98 to 100 percent across the domains, there were differences at the exact agreement level. Exact agreements for the model developed using the ELL training set ranged from 50 to 64 percent, while the blended training set resulted in exact agreements ranging from 66 to 76 percent. Pearson correlations for the two models were very similar (.83 to .89 for the first and .84 to .90 for the second). This study supports the use of a blended training set.

Background and Overview

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. 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.

Most of these studies have been conducted on essays written by students within the United States and have not targeted a particular group. This study targets Taiwanese English Language Learners in an effort to determine the optimal training set composition for accurately scoring essays submitted by these students. Before providing the details of the investigation, it is important to have a foundational understanding of how IntelliMetric™ works and why the training set plays such a critical role in the process.

How Does IntelliMetric™ Score Essays?

IntelliMetric™ is an intelligent scoring system that emulates the process carried out by human scorers (Elliot, 2002). 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.

The next section of this paper presents a recent study regarding the training set composition and its effects on scoring essays written by Taiwanese English Language Learners. This study investigates the quality of a training set that is targeted specifically for one particular group: Taiwanese English Language Learners. The research question posed is whether a training set composed entirely of Taiwanese ELL writers will provide a better IntelliMetric™ model than a model that also includes native speakers' essays within the training set.

The Investigation

The purpose of this investigation was to determine the effect of the inclusion of native English speakers' essays in the training set for an IntelliMetric™ model that would be used to score essays written by Taiwanese ELL students at the high school or college level. Previous studies have shown that IntelliMetric™ models score most accurately when the training set is representative of the population that will be writing to the model with respect to background, age, and range of general writing ability. For instance, a scoring model that was normed with elementary-level writers would not be a valid assessment for high school students.

This study investigates the training set composition as it affects the scoring accuracy for ELL writers. Specifically, the research question is: Do native English essays increase the accuracy of an IntelliMetric™ model for scoring essays written by Taiwanese English Language Learners?

Procedures

This study was conducted to determine the effect of the inclusion of native English speakers' essays in a training set developed to score essays written in English by students whose native language is Taiwanese.

Data Source. The data used as a basis of this research was collected through MY Access!®, an online writing instructional tool developed by Vantage Learning. Students in Taiwan submitted essays written to a particular pilot (field test) prompt available in MY Access!. Students in the United States who are native English speakers also submitted essays to the same prompt. Two expert human scorers scored each essay holistically and across five domains of writing: Focus and Meaning, Content and Development, Organization, Language Use and Style, and Mechanics and Conventions. Approximately 500 essays with scores from two raters on a six-point rubric were used in this study. The same six-point rubric was applied to all essays. Two hundred and fifty essays were submitted by Taiwanese students and 150 essays were submitted by native speakers. The composition and distribution of native speakers' essays were insufficient to build a Native Only model for this study. This will be investigated upon collection of additional data.

Data Preparation. The training set data were cleaned and prepared for IntelliMetric™ training. The training data were split into sets: one training set that included only Taiwanese essays and another set that included all of the Taiwanese essays plus the native English essays. Within each training set, the same set of 50 Taiwanese essays was kept blind for validation purposes.

The analyses were conducted "blind" to avoid the pitfall encountered in some essay scoring validation studies where the training and prediction are carried out on the same data set. A failure to separate training and validation artificially inflates results and contributes to false expectations for performance under operational conditions. In other words, the set of 50 validation responses was treated as unknown, while the second training set was used as a basis for "training" the IntelliMetric™ system.

Scoring. Following the creation of the IntelliMetric™ holistic and domain scoring models based on the training sets, the additional set of validation papers was scored by IntelliMetric. In addition, all essays in the training sets were scored by IntelliMetric.

Analysis. Following these initial steps, analyses were conducted to compare the expert and IntelliMetric™ scoring within the 50 validation papers withheld from the training sets in both scenarios. The analyses included a comparison of means, tabulation of agreement rates, and calculation of Pearson correlations.

<u>Results</u>

Comparisons of Means. The means and standard deviations for the expert and IntelliMetric™ scores were comparable (see **Table 1**). There were no significant differences revealed in the t-test of the means (p>.05). These data are summarized in the table below. Due to the use of the same set of 50 papers for validation, the human score data is the same in both scenarios.

Table 1: Descriptive Statistics

		Taiwanese ELL-Only Model		Taiwanese ELL and Native English Combined Model	
		Mean Score	Standard Deviation	Mean Score	Standard Deviation
Holistic	Human	3.54	1.25	3.54	1.25
	IntelliMetric	3.56	1.19	3.58	1.2
Focus and Meaning	Human	3.72	1.18	3.72	1.18
	IntelliMetric	3.36	1.3	3.7	1.14
Content and Development	Human	3.36	1.3	3.36	1.3
	IntelliMetric	3.4	1.2	3.4	1.2
Organization	Human	3.38	1.2	3.38	1.2
	IntelliMetric	3.22	1.01	3.4	1
Language Use and Style	Human	3.44	1.08	3.44	1.08
	IntelliMetric	3.26	1	3.5	1
Mechanics and Conventions	Human	3.68	1.05	3.68	1.05
	IntelliMetric	3.44	1.04	3.66	0.9

While there were no significant differences in means in either scenario, the average difference in mean is smaller for the blended training set than for the ELL-only training set.

Agreement Analysis. The frequency with which IntelliMetric™ was in agreement with scores assigned by expert graders was calculated to determine the extent to which IntelliMetric™ would yield scores similar to those identified by human scorers in practice. The extent to which IntelliMetric™ and expert scorers agreed exactly, were within one point of each other (adjacent agreement), or were two or more points apart (discrepant agreement) were calculated. **Table 2** shows the counts of exact matches, adjacent matches, and discrepant scores under each scenario.

Table 2. Agreement Rates

		Exact	Adjacent	Discrepant
Holistic	Taiwanese ELL-Only Model	82%	18%	0%
	Blended Model	80%	20%	0%
Focus and Meaning	Taiwanese ELL-Only Model	50%	48%	2%
	Blended Model	76%	22%	2%
Content and Development	Taiwanese ELL-Only Model	64%	36%	0%
	Blended Model	68%	32%	0%
Organization	Taiwanese ELL-Only Model	56%	44%	0%
	Blended Model	66%	34%	0%
Language Use and Style	Taiwanese ELL-Only Model	50%	50%	0%
	Blended Model	66%	34%	0%
Mechanics and Conventions	Taiwanese ELL-Only Model	62%	36%	2%
	Blended Model	76%	22%	2%

For each domain, the model based on the blended native and ELL training set produced higher exact agreement rates than the model based on the ELL-only training set.

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. The Pearson r correlation theoretically varies from –1 to +1. However, the true value of this statistic under operational conditions is highly dependent on the variance in the data set. Reduced variance will significantly underestimate the correlation.

Table 3. Pearson Correlations

	Taiwanese ELL-Only Model	Taiwanese ELL and Native English Model
Holistic	.94	.94
Focus and Meaning	.87	.89
Content and Development	.89	.90
Organization	.85	.89
Language Use and Style	.78	.84
Mechanics and Conventions	.83	.85

For each domain, the blended model yielded a higher Pearson correlation. The two holistic models had a correlation of .94.

Summary

The findings of this study show that the inclusion of the native English speakers' essays yielded a slight improvement in the domain scoring accuracy of the IntelliMetric™ system to score essays written by Taiwanese English Language Learners. The scoring model developed with only Taiwanese essays also performed very well. The holistic scoring models were so close to each other in accuracy that either model would be acceptable for use in scoring new Taiwanese ELL essays written to this prompt.

The results of this study confirmed the findings of previous studies that indicated holistic scoring agreement rates are typically higher than domain-level agreement rates. The use of the blended training set produced better domain-level agreement rates than the all-ELL training set, yet the domain-level agreement rates were still much lower than the holistic agreement.

The expert scorers indicated that the essays written by the Taiwanese English Language Learners were on average much better than those written by the United States native English students. This may have had some effect on the results of this analysis. Additional studies could be conducted that include a larger variety of Taiwanese ELL students who are really struggling with the English language.

Future research is being targeted at additional compositions of training sets, including essays written by English Language Learners with various native languages as well as a "Native Only" model. In addition, these studies will be carried out at each grade level (elementary, middle, high school) in order to determine if the relationship changes depending on the population of writers. With this complete set of data, the best practices for automated essay scoring training set composition for ELL essay scoring can be determined.

References

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

Elliot, S. 2002. From Here to Validity. In *Automated Essay Scoring*, ed. M. Shermis and J. Burstein. New Jersey: Lawrence Erlbaum Associates.

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.