Research Summary

IntelliMetric™ Scoring Accuracy Across Genres and Grade Levels

Vantage Learning

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

Research Summary

IntelliMetric™ Scoring Accuracy Across Genres and Grade Levels

Introduction

This document summarizes data regarding the scoring accuracy of IntelliMetric™, Vantage Learning's automated essay scoring system. IntelliMetric™ has been found to be an effective tool for scoring essay-type, constructed response questions across K-12, higher education and professional training environments as well as within a variety of content areas and for a variety of assessment purposes.

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.

IntelliMetric™ is trained to score essays much the same way as expert human raters are trained. Experts are provided anchor papers specific to the prompt, are given scores to those papers, and are taught why each paper should receive a certain score. The human raters are given additional scored papers for training and are ultimately asked to score some papers on their own. If the human scoring is acceptable with regards to the standard, the human rater is then allowed to score new essays for that particular prompt. Similarly, IntelliMetric™ is trained using a set of essays which have already been scored. This training allows the scoring engine to recognize what discourse elements of an essay written to a specific prompt are desirable. The IntelliMetric™ engine learns what it means to be an essay earning each score point on the rubric. This training process is a prompt-specific process in which a unique training set is used and a resulting unique IntelliMetric™ model is developed. The IntelliMetric™ scores will be reliable and valid for legitimate essays submitted to that prompt. An essay that is not written to this prompt cannot be accurately scored using this prompt-specific model. For example, a model created for a persuasive prompt on why books should be banned would not be appropriate for scoring essays written to a narrative prompt which asks students to write a story about their experiences during the first day of school.

How the Accuracy of IntelliMetric™ is Evaluated

Before it can be approved for instructional use, each IntelliMetric™ model must be rigorously evaluated to certify its accuracy in scoring essays. During the training process, a portion of every essay set is withheld from the training set used to create each IntelliMetric™ scoring model. These validation responses are scored by IntelliMetric™ and compared to the human expert scores. This provides a true comparison of blind scoring by IntelliMetric™ compared to scores provided by an expert(s).

Using this validation set to evaluate the accuracy of IntelliMetric™, the means of the humans and the IntelliMetric™ model are compared. If they do not differ significantly based on results of a t-test for difference in means, the agreement rates are calculated. An exact agreement rate refers to the proportion of essays in which the human and IntelliMetric™ score were identical, while an adjacent agreement rate refers to the proportion of scores that were within one point of each other on a 6-point scale. Any scores of essays that are found to be discrepant (more than one point apart) are also noted and reviewed. Finally, the Pearson correlation between the scores is calculated. The higher the positive correlation between the two scores, which can range from 0 to 1, the more associated the data values are with each other.

In order for an IntelliMetric™ model to be approved for use in MY Access!®, exact agreement rates must typically be at least 70%, adjacent agreement rates must approach 100%, and Pearson correlations must be at least 0.80. If all three of these requirements are met, the model can be approved for use in scoring essays in MY Access!®. For IntelliMetric™ models used in other programs, other standards of acceptable models may apply. Due to the training process, the quality of the IntelliMetric™ model is directly related to the quality and composition of the training set.

This research summary contains two main sections:

1) Sampling of IntelliMetric™ model performance data (agreement rates and Pearson correlations) across genres at the Elementary, Middle, and High School levels

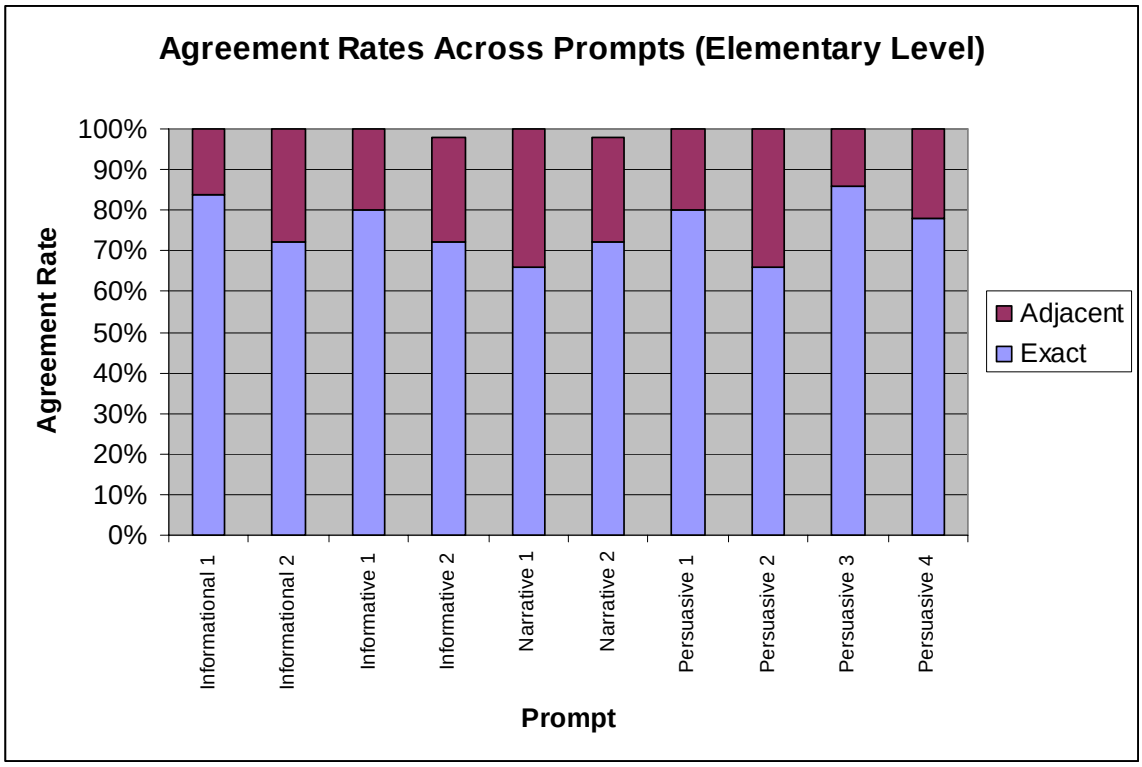
2) Comparison of Expert Scorer Agreement and IntelliMetric/Expert Agreement at the Higher Education Level

Sampling of IntelliMetric™ Model Performance Data

Upper Elementary Level (Grades 4-5)

Analysis of Agreement Rates. The agreement rates of ten different IntelliMetric™ models, across four different writing genres at the Upper Elementary Level, are shown in **Figure 1**. Exact agreement rates for these models ranged from 66% to 86%, with adjacent agreement rates of 100% for nearly every prompt. On average, the exact agreement rate across these ten prompts is 76% and the adjacent agreement rate is 100%.

Figure 1.



Calculation of Pearson Correlations. The Pearson correlations for these same ten prompts are shown in **Table 1**. The Pearson correlations ranged from 0.87 to 0.96, with an average Pearson correlation of 0.93 across the ten prompts.

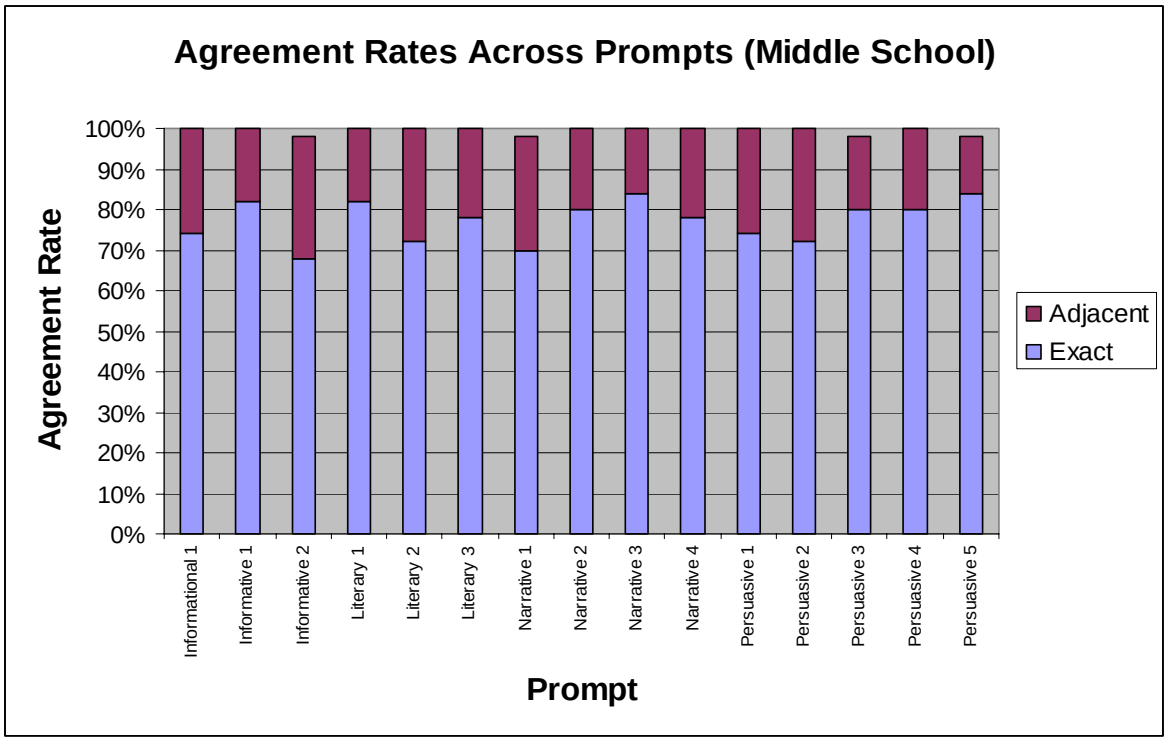
Table 1. **Elementary School Level Prompt Pearson Correlations Between Experts and IntelliMetric™**

Prompt	Pearson Correlation
Informational 1	0.96
Informational 2	0.92
Informative 1	0.95
Informative 2	0.91
Narrative 1	0.91
Narrative 2	0.87
Persuasive 1	0.95
Persuasive 2	0.91
Persuasive 3	0.96
Persuasive 4	0.95
Average	**0.93**

Middle School Level (Grades 6-8)

Analysis of Agreement Rates. The agreement rates of fifteen different IntelliMetric™ models, across five different writing genres at the Middle School Level, are shown in **Figure 2.** Exact agreement rates for these models ranged from 68% to 84%, with adjacent agreement rates of 100% for nearly every prompt. On average, the exact agreement rate across these fifteen prompts is 77% and the adjacent agreement rate is nearly 100%.

Figure 2.



Calculation of Pearson Correlations. The Pearson correlations for these same fifteen prompts are shown in **Table 2**. The Pearson correlations ranged from 0.87 to 0.96, with an average Pearson correlation of 0.93 across the fifteen prompts.

Table 2. Middle School Level Prompt Pearson Correlations Between Experts and IntelliMetric™

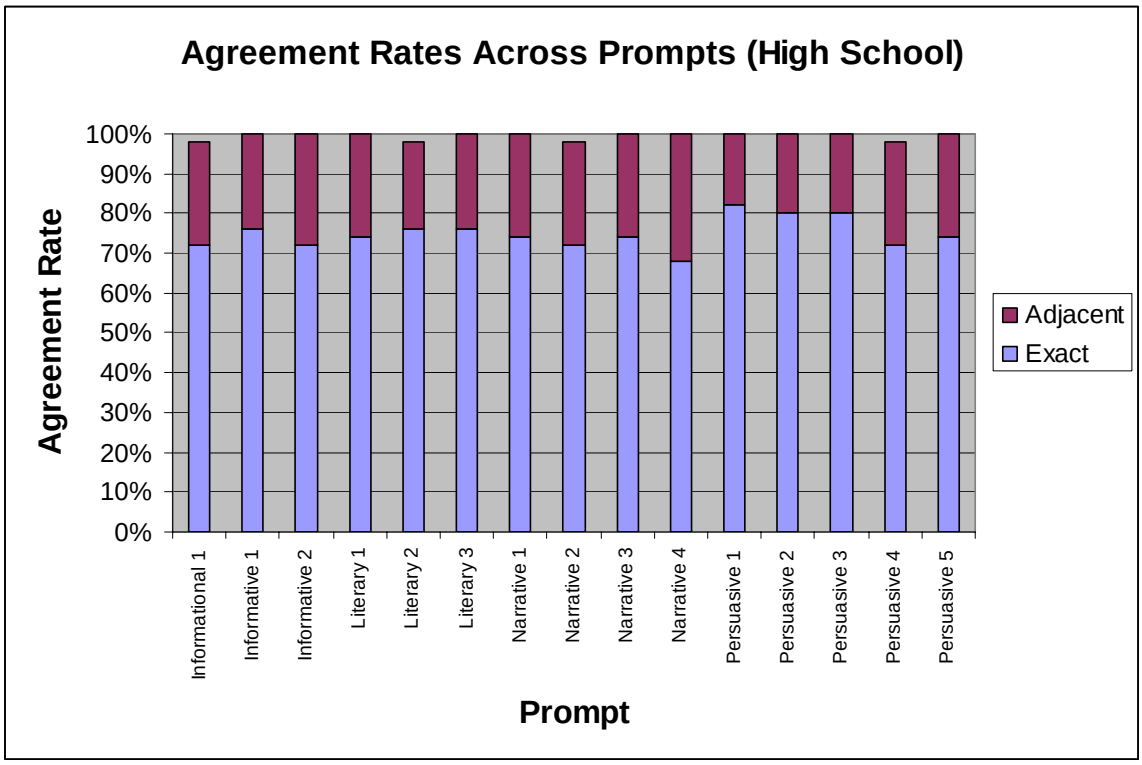
Prompt	Pearson Correlation
Informational 1	0.91
Informative 1	0.95
Informative 2	0.90
Literary 1	0.95
Literary 2	0.93
Literary 3	0.94
Narrative 1	0.86
Narrative 2	0.94

Narrative 3	0.96
Narrative 4	0.93
Persuasive 1	0.93
Persuasive 2	0.93
Persuasive 3	0.94
Persuasive 4	0.94
Persuasive 5	0.93
Average	**0.93**

High School Level (Grades 9-12)

Analysis of Agreement Rates. The agreement rates of fifteen different IntelliMetric™ models, across five different writing genres at the High School Level, are shown in **Figure 3**. Exact agreement rates for these models ranged from 68% to 82%, with adjacent agreement rates of 100% for nearly every prompt. On average, the exact agreement rate across these fifteen prompts is 75% and the adjacent agreement rate is nearly 100%.

Figure 3.



Calculation of Pearson Correlations. The Pearson correlations for these same fifteen prompts are shown in **Table 3**. The Pearson correlations ranged from 0.89 to 0.95, with an average Pearson correlation of 0.92 across the fifteen prompts.

Table 3. High School Level Prompt Pearson Correlations Between Experts and IntelliMetric™

Prompt	Pearson Correlation
Informational 1	0.91
Informative 1	0.94
Informative 2	0.91
Literary 1	0.91
Literary 2	0.92
Literary 3	0.92
Narrative 1	0.93
Narrative 2	0.90
Narrative 3	0.93
Narrative 4	0.92
Persuasive 1	0.95
Persuasive 2	0.95
Persuasive 3	0.94
Persuasive 4	0.89
Persuasive 5	0.93
Average	**0.92**

A Comparison of Expert Agreement and IntelliMetric™/Expert Agreement

This section of the summary details a separate investigation evaluating how often human experts agree with each other compared to how often IntelliMetric™ agrees with the experts. Three different persuasive validation sets of essays written by college students were used for analysis. These responses were scored by two expert scorers and IntelliMetric™. These essays were not used in the training set used to create the IntelliMetric™ model, therefore providing a true comparison between blind scoring by IntelliMetric™ and scores provided by experts.

Analysis of Agreement Rates. The agreement rates for three higher education persuasive prompts are shown in **Table 4**. The exact agreement rate between the two expert scorers ranged from 58% to 61% while the exact agreement rate between IntelliMetric™ and the final expert score (the average of the two expert raters) ranged from 65% to 70%. The adjacent agreement rate between the expert raters ranged form 96% to 99%, with as many as 4% discrepant essays, while the adjacent agreement between IntelliMetric™ and the final expert score ranged from 98% to 100%, with discrepant essays for only one of the three prompts.

Table 4. Comparison of Expert Agreement and Agreement Between IntelliMetric™ and the Experts

Prompt	Agreement Rate Between Two Expert Raters			Agreement Rate Between IntelliMetric™ Score and Final Expert Score		
	Exact	Adjacent	Discrepant	Exact	Adjacent	Discrepant
Persuasive 1	58%	96%	4%	69%	100%	0%
Persuasive 2	61%	99%	1%	70%	98%	2%
Persuasive 3	60%	96%	4%	65%	100%	0%

Calculation of Pearson Correlations. The Pearson correlations for these same three prompts are shown in **Table 5**. For all three prompts, the Pearson correlations between the IntelliMetric™ and final expert score are either equal to or higher than those between two expert raters.

Table 5. Pearson Correlations

Prompt	Pearson Correlation Between Two Expert Raters	Pearson Correlation Between IntelliMetric™ Score and Final Expert Score
Persuasive 1	0.77	0.85
Persuasive 2	0.80	0.80
Persuasive 3	0.72	0.80

Conclusion

The analysis shown confirms that IntelliMetric™ can reliably and accurately score student essays. Across all educational levels and writing genres, IntelliMetric™ is able to achieve a very high rate of agreement with expert scores. For all of the prompts used in this summary, on average, IntelliMetric™ achieved an exact agreement rate of 76% and an adjacent agreement rate greater than 99%. The majority of models achieved a 100% adjacent agreement rate. Finally, the Pearson correlations between the human and IntelliMetric™ scores averaged 0.93 across all forty prompts, indicating that the linear relationship between human scores and IntelliMetric™ scores was extremely strong. These findings attest to the overall quality of IntelliMetric™ in scoring essays.

In addition, it was found that IntelliMetric™ agreed with the experts more often than the experts agreed with each other. For all three prompts, the exact agreement rates were markedly higher between IntelliMetric™ and the final human expert score. These findings indicate that IntelliMetric™ is able to score essays with a level of accuracy that meets or exceeds the level obtained by two expert scorers.

The information contained in this summary supports the conclusions that IntelliMetric™:

1. Accurately scores open-ended responses across a variety of grade levels, subject areas and contexts
2. Agrees with expert scoring, often exceeding the performance of expert scorers
3. Shows stable results across samples.

