AES Data Report

Machine vs Human Scores

For

New Zealand Qualifications Authority

Prepared by

New Data Solutions Pty Ltd and Vantage

27 July 2021

Contents

Project Overview 3
Project Summary 3
Report on the analysis of Blind Scores (Validation Set) 5
Standard Scripts Summary 5
Standard Scripts Recommendations: 6
Standard Script Blind (Validation) Results 7
Non-Standard Results 8
APPENDIX A —Preliminary Data Report 10
APPENDIX B —Standard Scripts Blinds Scores 14
APPENDIX C – How IntelliMetric Works 14
About IntelliMetric ™ 15
What IntelliMetric™ cannot do 16
How do we know IntelliMetric works? 22

Project Overview

In May of 2021 New Data Solutions Pty Ltd (NDS) was awarded the contract to undertake further analysis of the suitability of automated marking, specifically using IntelliMetric™, for the writing component of NZQA exams. NZQA is investigating the reliability and efficiency that automated marking might provide as well as undertaking a more robust review of how automated essay scoring handles non-scoreable papers.

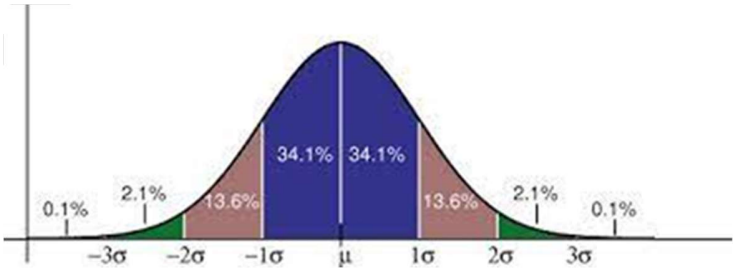
The preliminary report provided an overview of how IntelliMetric works followed by an explanation of the steps taken to that point and recommendations for improvement. The report included the data from our sub-set cross validations for each model including a confusion matrix, Pearson, Kern and Cohen's kappa indices. This report is included at Appendix A.

This report provides a summary analysis of approximately 500 papers for each of four prompts with Exact and Adjacent Match, Discrepant and Super-Discrepant (>+2) as well as the Quadratic Weighted Kappas and Matthews Correlation Coefficient, the now popular metric for AI and Machine Learning.

Project Summary

In order to create a Gold Standard IntelliMetric Model, the engine requires a minimum of 25-30 scripts (exemplars) at the tails of the score scale and a roughly normal distribution across the remaining score points. Anchoring the tails is essential to allow the engine to understand the upper and lower boundaries of the range. The greater quantity of scripts is required in the middle-points of the scale, the peak of the normal distribution, because these scripts tend to be the hardest to differentiate at these score points. When thinking of the engine in this regard it is helpful to think of humans looking at scripts at the low, middle and high-end of the scale. The low scoring scripts and the high scoring scripts are often so clearly exemplary of poor quality or very good quality, that humans will easily agree on the scores. The papers in the middle range require a greater attention to detail and are often harder to differentiate between scores of 4 and 5 because the differences are less clear.

The ideal distribution for NZQA training scripts should attempt to match the following score distribution:



The data presented to NDS/Vantage for modeling was grossly insufficient for creating a Gold Standard model for any of the four prompts that were modeled. The table below show the training set counts by prompt, by score:

	Count of Scores by Prompt			
Score	91005Q1	90849Q3	90850Q1	90850Q3
0	1	5	2	2
1	21	18	9	7
2	25	30	30	30
3	30	30	30	30
4	30	30	30	30
5	30	29	30	30
6	30	30	30	30
7	30	30	30	30
8	30	30	15	30

As can be seen in the counts above most of the prompts included a flat unimodal distribution. Additionally, all four prompts were below standard for the score point of 1 and 90850Q1 was also insufficient at the high-end of the scale. The papers listed as score point 0 (zero) were not included in the training. Later in the report we discuss the recommendations related to these counts.

Despite the training data, IntelliMetric was able to create relatively strong models and identify papers that should be reviewed by humans as they triggered one or more of the non-standard flags provided by IntelliMetric. Three of the four models could reliably be operationalised today and all four models could be vastly improved through remodeling with an appropriate distribution.

We also see several areas for improvement beyond the representative nature of the data set and the scale of the rubric.

Report on the analysis of Blind Scores (Validation Set)

Standard Scripts Summary

Early in week commencing, 2 July, the NDS/Vantage team received a large data set from NZQA consisting of approximately 500 scripts per prompt. These had been scored and the scores were withheld from NDS/Vantage. This data was to be used to validate the IntelliMetric models created in the weeks prior for four unique prompts. (Additional information on the initial models and preliminary report is available at the end of this document.)

This validation data, scored-blind scripts, were shared via a secure file sharing folder in MS-Excel spreadsheet format. ***NO SCORES*** for any marking criteria were included in the data to ensure the scores provided were generated by the IntelliMetric engine.

The NDS and Vantage team invested time in extracting and cleaning the files and preparing them to be ingested by IntelliMetric. During this process extreme diligence was used to ensure the text remained true to what the human markers would have seen. IntelliMetric focuses on the content and linguistic elements of the writing and, as such, non-text elements and formats impact the engine's ability to read the files. The cleaning included:

- Extraction of the script ID and associated student response from Excel to .txt file and editing to the IntelliMetric format
- Removal of UTF encoding, tabs etc

Once prepared the data was sent to the appropriate Intellimetric model. Scores for each of scripts were generated. Additionally, IntelliMetric applied non-standard flags for many of the responses. Thresholds were set low for the various non-standard flags and it is recommended that these be revisited and possibly adjusted by NZQA following a more thorough analysis. The scores and non-standard flags were shared with NZQA.

On July 24th NZQA shared the scores for the score-blind papers with NDS/Vantage. Scores included a single score for each script as marked by an individual Human Marker. The NDS/Vantage team shared the IntelliMetric marks and the Human Marks with the New Zealand Council For Educational Research (NZCER) who proceeded to undertake an analysis of the M1 to machine scores.

The following statistics were generated:

- Exact Match (Marker 1 and Marker 2 match exactly)
- Adjacent Match (Marker 1 and Marker 2 assign adjacent scores)
- Discrepant Match (Marker 1 and Marker 2 assign scores off by 2)
- Super Discrepant Match (Marker 1 and Marker 2 assign scores off by >2)
- Pearson Correlation
- Cohen's Kappa
- Quadratic Weighted Kappa
- Mathews Correlation Coefficient
- Mean and Standard Deviation for both Human and Machine Scores
- Effect Size

Three of the four models meet the threshold required to be operationalized today. The fourth model, while close to an operational standard should be revisited. We believe all four models would dramatically improve with an appropriate distribution of training papers.

Standard Scripts Recommendations:

- Provision of additional training papers at the tails of the score scale such that we meet the minimum of 25-30 papers at the tails. The absence of sufficient papers at the tails prevents IntelliMetric from learning the full set of features humans would use to define the score points in those ranges, making it challenging for IntelliMetric to anchor that criteria and specifically define the tails. (See Appendix A for counts of responses per prompt.)
- Provision of papers to meet a normal distribution, or substantially closer to normal, inclusive of the 30 papers at the tails.
- Review the general selection of training papers for their representative nature (i.e. only include papers which would be used for human rater training, eliminating papers of a single word etc.)
- Provide clean asci text files or move to data transmission via the WriteShift API to eliminate the possibility of human error during the editing process.
- Review the inclusion of score point 0 papers and ensure the appropriate scale is used.
- Review how humans would compare to one-another in order to understand machine performance. Indications in the data, notably the standard deviation suggest the engine may have been more reliable than humans.
- Review the decimal values provided by the machine to human marks to identify issues with the human scoring or potential issues with the rubric.

Standard Script Blind (Validation) Results

Below is the New Zealand Council for Educational Research (NZCER) data analysis with commentary by the NDS/Vantage team. NZCER analysis was undertaken by Dr. Elliot Lawes and presented to NDS/Vantage on 26 July 2021. Under contract to NDS, NZCER were given scoring data for NZQA essays 90849Q3, 90850Q1, 90850Q3, and 91005Q1.

For each of these essays this scoring data consisted of matched lists of Intellimetric machine scores and NZQA human marker scores for approximately 500 sample essays. NZCER then independently carried out a number of analyses to assess the similarity of the machine scores and human scores.

The analyses were those commonly used by NDS and Vantage. They are summarised in the table of results below followed by commentary by NDS/Vantage.

Essay	Exact	Adj	Exact+Adj	Off2	SDsc	Pearson	Kappa	QuadWKappa	Matthews
90849Q3	0.360	0.413	0.773	0.178	0.050	0.663	0.215	0.635	0.217
90850Q1	0.463	0.423	0.886	0.091	0.023	0.764	0.334	0.762	0.337
90850Q3	0.368	0.438	0.807	0.154	0.039	0.714	0.236	0.702	0.238
91005Q1	0.389	0.460	0.849	0.124	0.027	0.744	0.266	0.744	0.269

For the four prompts above the statistical measures suggest that the IntelliMetric engine was able to understand the scoring. Prompts 50Q1, 50Q3 and 05Q1 have sufficiently high Pearson and Quadratic Weighted Kappa to suggest they are ready for operational scoring based on our experience with thousands of models similar to these. We would recommend additional papers both at score point 1 and sufficient papers to create a normal distribution across other score points.

Based on our experience, we recommend prompt 49Q3 receive re-training with additional training papers to ensure reliability is better than humans.

Essay	Nraters	mean_x	sd_x	mean_y	sd_y	effect_size
90849Q3	484	4.362	1.387	4.756	1.557	-0.268
90850Q1	482	3.950	1.407	4.048	1.476	-0.068
90850Q3	486	4.681	1.551	4.959	1.551	-0.179
91005Q1	483	5.068	1.578	5.097	1.577	-0.018

In this table, mean_x and sd_x are the mean and standard deviation of the human scores respectively, and mean_y and sd_y are the mean and standard deviation of the machine scores. Therefore, a negative effect size indicates that human scores tended to be lower than machine scores. This data suggests the human variability may be impacting the overall models. The standard deviations for 05Q1, 50Q3 and 05Q1 indicate the engine is locked in on the scoring and the variability is therefore attributed to the humans.

Non-Standard Results

A key component of any large-scale, automated scoring system is the ability to identify and classify non-standard scripts. Identifying these scripts can mitigate both false-positive and false-negative reporting and will build confidence in the overall marking system.

Identifying non-standard scripts may be easy in some instances, particularly those involving a response consisting entirely of emojis or scripts that are blank. However, many times these determinations are often considered unquantifiable or judgement calls and thus a uniquely human capability. As such, the challenges of quantifying the elements that define a unique response versus one that is truly non-standard must be handled by the automated scoring system in a humanistic way. Through advanced Artificial Intelligence and tuning of the engine it can be adjusted to match the judgement of the human markers for each unique setting and establish mutually agreeable thresholds.

Through the use of our proprietary LegitiMatch™ technology embedded within IntelliMetric, responses that appear off topic, are too short to score reliably, do not conform to the expectations for edited Australian/New Zealand English or are otherwise unusual are identified as part of the training process.

In the table below we identify the possible Non-Standard identifiers currently in use by the IntelliMetric AES system and returned via the WriteShift API for operational scoring.

Non-Standard Code	**Descriptor**
OK	Scores returned with no issues
Too Short	Too Short to grade
Off Topic	Off-topic
Repetitious	Repetitious
Insufficient Development	Inadequate Development
Unknown Words	Contains too many unknown words
Bad Syntax	Contain major syntax problems
Violent	Contains unnecessary violent language
Copied the Question	Substantially copied the question

Each of the categories above should be thought of in a humanistic way. There are, however, two primary differences. The engine will trend towards greater reliability and greater precision in the identification of non-standard scripts and thus the engine can be fine-tuned to meet the requirements of NZQA. Using guidance provided by NZQA psychometricians and writing experts we would anticipate tuning the non-standard thresholds to meet your requirements prior to operational scoring. For the initial non-standard analysis all thresholds were set to exceedingly liberal levels, providing scores where, upon tuning, they would be identified as non-standard.

The table below identifies the scripts identified by IntelliMetric as non-standard and a reason for the classification. In an operational setting we recommend working with Vantage to tune this appropriately for your project and when a paper is classified as non-scorable route them for human review and confirmation of the non-standard code or scores.

4923467844	Major Syntax		4923001263	Off-Topic		4923880294	Major Syntax	4918799240	Major Syntax
4926026377	Major Syntax		4923879708	Major Syntax		4007261388	Repetitious	4007019524	Off Topic
3994925735	Major Syntax		4928733888	Off-Topic		4924159622	Off Topic	4002420870	Gobbledegook
4923692997	Not Enough Info		4926023956	Off-Topic		4928248955	Repetitious	4002420870	Major Syntax
4938437013	Major Syntax		4925817898	Major Syntax		4930319015	Major Syntax	4928484790	Off Topic
4927514746	Major Syntax		4096547819	Repetitious		4931239582	Major Syntax	4925808177	Repetitious
4928616510	Major Syntax		4924188629	Major Syntax		4926726939	Off Topic	4928556469	Repetitious
4928237831	Major Syntax		4923576977	Repetitious		4926603476	Off Topic	4935719146	Gobbledegook
4923503641	Major Syntax		4924847343	Repetitious		4927328130	Off Topic	4935719146	Major Syntax
4925814556	Major Syntax		4210415276	Off-Topic		4923003381	Major Syntax	4925972335	Major Syntax
4927318136	Major Syntax		4925827949	Major Syntax		4001440920	Major Syntax	4925477029	Off Topic
4930343132	Major Syntax		4920158606	Off-Topic		4164640061	Major Syntax	4918798975	Major Syntax
4937815988	Major Syntax		4921823524	Repetitious		4024906990	Major Syntax	4919895515	Major Syntax
4930464571	Major Syntax		4928854873	Off-Topic		4923713208	Off Topic	4571809453	Repetitious
4521520593	Off Topic		4921808449	Major Syntax				4918796675	Major Syntax
			4929397703	Off-Topic				4918801422	Major Syntax
			4922925349	Off-Topic				4935718961	Major Syntax
			4925834574	Off-Topic				4938445920	Major Syntax
								4922960800	Major Syntax

APPENDIX A —Preliminary Data Report

On Friday 16 July 2021 the NDS/Vantage team received data from NZQA to train IntelliMetric to score 4 prompts. This data was shared via secure file sharing folder in a MS-Excel spreadsheet that included standard, unique identifiers, student responses and scores for each paper.

The NDS and Vantage team invested time to clean and prepare the files for the IntelliMetric modeling process. IntelliMetric focuses on the content and linguistic elements of the writing and, as such, non-text elements and formats impact the engine's ability to read the files. The cleaning included:

- Extraction of the student ID and associated student response from Excel to .txt file and editing to the IntelliMetric format
- Removal of UNIX encoding
- Extraction of student ID and associated scores into .csv format

From this data, four IntelliMetric models were created; one for each prompt. As part of this process, student samples were held back from each model to be used as a sub-set cross validation. It is standard practice when creating a model in IntelliMetric to withhold some scripts with known scores from the training set. IntelliMetric uses these scripts for some gross adjustments to the engine to determine if the model is performing well or not.

The first analysis of the data indicated strong Pearsons for most of the prompts. The strong Pearsons combined with the sub-standard data set suggests that with some minor adjustments these models can be greatly improved. We are confident the models can be improved with further collaboration.

Recommendations:

- Provision of additional papers at the tails of the score scale such that we meet the minimum of 30 papers at the tails. The absence of sufficient papers at the tails prevents IntelliMetric from learning the full set of features humans would use to define the score points in those ranges. Some criteria have fewer than 10 responses at the tails making it challenging for IntelliMetric to anchor that score point and specifically define the tails (see Appendix A for counts of responses per prompt)
- Review the general selection of training papers for their representative nature
- Provide clean asci text files to eliminate the possibility of human error during the editing process
- Start the score scale at 1 rather than zero.

Preliminary Model Results

For each prompt below we have provided a confusion matrix indicating where the computer and human matched, were adjacent, were discrepant or were super-discrepant (off by 2+).

Below each confusion matrix we have provided a count of the exact, adjacent, discrepant and super-discrepants. Ideally the confusion matrix will show a strong correlation from top left to bottom right. This analysis helps us understand if the models are deployable and where weaknesses may appear in the models. We also see prevalence issues from the training sets appear in the confusion matrix. In addition, we calculated a Pearson Correlation, Kern Index and Cohen's Kappa.

The next phase of the project involves scoring the "blind papers" for NZQA. We will provide both decimal and integer scores for these papers. This often provides further insights related to the machine vs human scoring as well as identifying where humans may have drifted from the rubric or training.

Scripts were shared with NDS/Vantage in a Microsoft Excel file. This format is not compliant with our recommended format and requires conversion to ASCI format for ingestion into IntelliMetric. We also note that several of the papers seemed exceedingly long, surpassing 10,000 characters which is notable for writing at this level.

The training sets and ultimately the models would likely improve if NZQA could:

- Remove all zero scores as those were ignored in the training process focusing only on scores 1-8
- Provide a minimum of 300 training papers
- Provide a minimum of 25-30 papers at the tails or the low and high-points in the rubric
- Provide a roughly normal distribution of papers across the score points

All models had reasonably good or good Pearson's Correlation and we recommend moving forward with the scoring of the blind papers.

Prompt: 90849Q3

Name	C1	C2	C3	C4	C5	C6	C7	C8
H1	2	3	0	0	0	0	0	0
H2	1	1	4	0	3	0	0	0
H3	0	0	2	2	1	0	1	0
H4	0	0	3	2	0	1	0	0
H5	0	0	1	2	3	0	0	0
H6	0	0	0	2	0	4	0	0
H7	0	0	0	1	0	1	1	3
H8	0	0	0	0	0	1	3	2

Exa 17 Adj 22 Off2 6 SDsc 5 Pear:0.795 (K):-0.100 Kap:0.247

Prompt: 90850Q1

Name	C1	C2	C3	C4	C5	C6	C7	C8
H1	0	2	0	0	0	0	0	0
H2	1	4	4	1	0	0	0	0
H3	0	2	3	1	1	0	0	0
H4	0	0	1	4	2	0	0	0
H5	0	0	2	0	3	2	0	0
H6	0	0	0	2	2	1	1	1
H7	0	0	0	0	2	3	1	1
H8	0	0	0	0	1	0	1	1

Exa 17 Adj 23 Off2 9 SDsc 1 Pear:0.819 (K):-0.060 Kap:0.230

Prompt: 91005Q1

Name	C1	C2	C3	C4	C5	C6	C7	C8
H1	4	0	0	1	0	0	0	0
H2	3	0	2	0	1	0	0	0
H3	1	1	2	0	0	0	0	0
H4	0	0	1	2	2	2	0	0
H5	0	0	1	2	1	1	2	0
H6	0	0	0	0	0	4	2	1
H7	0	0	0	0	2	2	3	0
H8	0	0	0	0	0	1	5	1

Exa 17 Adj 21 Off2 10 SDsc 2 Pear:0.840 (K):-0.140 Kap:0.245

Prompt: 90850Q3

Name	C1	C2	C3	C4	C5	C6	C7	C8
H1	0	1	1	0	0	0	0	0
H2	0	3	2	0	0	1	0	0
H3	0	1	0	3	3	0	0	0
H4	0	0	2	2	3	0	0	0
H5	0	0	0	3	2	2	0	0
H6	0	0	0	1	2	1	3	0
H7	0	0	0	0	1	1	2	3
H8	0	0	0	0	1	1	4	1

Exa 11 Adj 30 Off2 7 SDsc 2 Pear:0.791 (K):-0.140 Kap:0.095

Count of Scores Per Prompt (Training Set)

Red highlight = less than recommended number of scripts

	Count of Scores by Prompt			
Score	91005Q1	90849Q3	90850Q1	90850Q3
0	1	5	2	2
1	21	18	9	7
2	25	30	30	30
3	30	30	30	30
4	30	30	30	30
5	30	29	30	30
6	30	30	30	30
7	30	30	30	30
8	30	30	15	30

APPENDIX B —Standard Scripts Blinds Scores

Provided as at attachment via EXCEL

APPENDIX C – How IntelliMetric Works

How does IntelliMetric® Score Essay responses?

Evaluating examinee skills based on a written assessment is certainly not a new phenomenon. Accounts of written assessments date back several hundred years B.C. in the Chinese Civil Service System. While we may no longer lock the examinees in a prison-like setting refusing release until they have completed the assessment as the Chinese once did, today's writing assessments bare more similarity to ancient Chinese civil service testing than we care to admit. Still, written assessments have undergone some changes over the centuries.

Arguably, one of the most notable innovations in written assessment is the advent of automated essay scoring, or the use of computers to assist in the evaluation of written responses to assessment questions. The automated essay scoring movement dates back to the early 1960's. In the early to mid 1960's Dr. Ellis Paige demonstrated that a computer can be used to score student written responses to essay questions. Automated essay scoring has come a long way since then, however, Dr. Paige still deserves recognition and credit for the earliest practicable automated essay scoring system. His vision and innovation gave birth to today's automated essay scoring systems.

Rolling the clock forward a few decades, Vantage Learning's IntelliMetric™ automated essay scoring system has taken the reins by defining the state of the art in automated essay scoring. IntelliMetric is based on research and development stemming back to the 1980's and has been used successfully to score open-ended essay-type assessments since 1998. IntelliMetric™ was the first commercially successful tool able to administer and mark open-ended questions and provide immediate feedback to students in a matter of seconds.

With the growing interest in automated essay scoring have come many questions. This paper hopefully provides answers to some of those questions.

Introduction

Computers are everywhere. Their presence can be felt in almost every facet of our lives. From the workplace to the home, computers have taken on new roles. Places, that were computer-free only a few short years ago, now depend on them for day-to-day operations. We depend on computers every time we make a telephone call, drive our car, or make a transaction at the bank.

It is no surprise then that computers have become pervasive in education as well. From the student desktop to the administrative corridors, the presence of computers can be felt. Most recently computers have taken on a major role in educational assessment. Assessments at the national, state, district and classroom level are increasingly being delivered via computer. One computer application that has become quite important in education is the scoring of student responses to open ended questions. IntelliMetric™ is the most widely used system for scoring open-ended assessments.

About IntelliMetric ™

According to Elliot (2002) IntelliMetric™ is an intelligent scoring system that emulates the process carried out by human scorers. IntelliMetric is theoretically grounded in the traditions of Cognitive Processing, Computational Linguistics and Machine Learning. 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 the pooled judgments of the human scorers. Relying on Vantage Learning's proprietary CogniSearch™ and Quantum Reasoning™ Technologies, the IntelliMetric™ system internalizes the characteristics of the responses associated with each score point and applies this intelligence in subsequent scoring.

IntelliMetric works a lot like the holistic scoring systems commonly employed to score large-scale writing assessments. A group of individuals asked to score essay papers are provided with examples of each score point determined by experts. After internalizing the characteristics associated with each score point and demonstrating calibration with the expert-assigned scores, the group is asked to score the remaining papers whose scores are unknown. Much like human scorers who are generally trained on each specific question or prompt, IntelliMetric™ creates a unique solution for each prompt. This process leads to high levels of agreement between the scores assigned by IntelliMetric™ and those assigned by human scorers.

IntelliMetric is based on a blend of artificial intelligence, natural language processing and statistical technologies. IntelliMetric learns the characteristics of the score scale through exposure to examples of essay responses previously scored by experts. In essence, IntelliMetric internalizes the pooled wisdom of many expert scorers.

IntelliMetric uses a multi-stage process to evaluate responses. First, IntelliMetric is exposed to a subset of responses with known scores from which it derives knowledge of the scoring scale and the characteristics associated with each score point. Second, the model reflecting the knowledge derived is tested against a smaller set of responses with known scores to validate the model developed. Third, after making sure that the model is scoring as expected, the model is applied to score novel responses with unknown scores. Using Vantage Learning's proprietary Legitimatch™ technology, responses that appear off topic, are too short to score reliably, do not conform to the expectations for edited American English or are otherwise unusual are identified as part of the process.

IntelliMetric can be used for standardized assessments where a single essay submission is required as well as various instructional applications where a student can provide multiple submissions of an essay response and receive frequent feedback. IntelliMetric also offers various editing and revision tools such as a spell checker, grammar checker, dictionary, and thesaurus. The IntelliMetric tool provides feedback on overall performance, diagnostic feedback on several rhetorical and analytical dimensions of writing (e.g., conventions, organization), and provides detailed diagnostic sentence-by-sentence feedback on grammar, usage, spelling and conventions.

Gaining Acceptance. People often fear and misunderstand new technologies, particularly those that automate some element of human activity. Throughout history, people have feared and resisted technologies that insert themselves into activities previously reserved for humans. From the Luddite resistance to the automation of looms in England centuries ago to modern day resistance to the

automobile, there is no lack of examples of fear of technology. Automated essay scoring is certainly no exception.

The evaluation of student written work has been the purview of humans since the birth of the written word. So it comes as no surprise that the introduction of computers into this mix would raise a few eyebrows. But, as with most new technologies, a better understanding of what IntelliMetric™ is and what it is not often helps to erase these fears.

IntelliMetric is in good company. While the promise of artificial intelligence has not been fully met, many applications, based on the same principles as IntelliMetric, have been successful. For example, since the 1960's the academic community has explored the use of computers to help with medical diagnoses. Computers programmed based on the experience of experts can be consulted to make effective diagnoses for novel cases.

What IntelliMetric™ cannot do

As impressive as IntelliMetric is, it does have some limitations. Before turning to an explanation of how IntelliMetric works, let us take a few moments to talk about what IntelliMetric does ***not*** do.

IntelliMetric cannot think in the traditional sense of this word. Unfortunately (or fortunately depending on your perspective) the human brain is far more sophisticated than IntelliMetric can ever hope to be. IntelliMetric cannot independently score essays without significant input from experts. It is merely a tool (albeit a sophisticated one) for applying the thinking of experts to novel situations—information gained from known-score essays is applied to unknown essays. In short, while IntelliMetric seeks to model a human brain to score essays, it pales in comparison to the human brain.

IntelliMetric™ is far from infallible. It can and does make mistakes Still, it makes fewer errors than do human scorers. Interestingly, while critics of automated scoring are quick to point this out, human scoring may be subjected to far less scrutiny. Unfortunately, any process is fallible, whether undertaken by humans or computers.

Finally, IntelliMetric™ is not magic black box. It is not a mysterious unknown force. It is the product of established scientific principles which are both explainable and repeatable. While looking for the gears and detailed mechanisms powering IntelliMetric™ is unlikely to bear fruit, there is a clear set of processes, well-grounded in theory, that drive IntelliMetric. These are described below.

Even with these limitations in mind, IntelliMetric is still more successful at scoring responses to essay questions than are most human scorers. IntelliMetric compensates for its limitations in three key ways:

1. **IntelliMetric consistently applies the internalized rubric**. Once IntelliMetric learns the rubric and standards for scoring it never waivers from that rubric. Human scorers are notorious for having difficulty "sticking with" the rubric. A cup of coffee or a rest break can lead to a drift in criteria and standards and it is very difficult for a human marker to score the first and last paper in a set exactly the same way. IntelliMetric on the other hand can maintain the exact same standards throughout the process.
2. **IntelliMetric scores consistently over time**. IntelliMetric will produce the same scores for a given response from time to time. If IntelliMetric assigns a score of "1" today, it will continue to do so tomorrow, the day after, etc., ad infinitum. The same cannot be said for human scorers.