VANTAGE

www.vantagelearning.com

RB-929

How Does IntelliMetric^{TM} Score Essay Responses?

VANTAGE LEARNING

VANTAGE LEARNING
Measuring Success One Student At A Time.

RB-929
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 to release them until they have completed the assessment like 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 1960s. In the early to mid-1960s Dr. Ellis Paige demonstrated that a computer could be used to score student written responses to essay questions. Automated essay scoring has come a long way since its infancy in the 1960s, but 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 1980s and has been used successfully to score open-ended essay-type assessments since 1998. IntelliMetric was the first commercially successful tool able to administer 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 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 computers. We depend on computers every time we make a telephone call, drive our car or make a transaction at the bank.

It comes as no surprise 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 on 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

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'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'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 inserted 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 raises a few eyebrows. But, as with most new technologies, a better understanding of the technology can help. By understanding what IntelliMetric is and what it is not can help 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 1960s, 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. It is not a mysterious unknown force. It is the product of established scientific principles that are both explainable and repeatable. While looking for the gears and detailed mechanisms powering IntelliMetric is unlikely to be fruitful, there is a clear set of processes, well-grounded in theory, that do drive IntelliMetric.

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.

- **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; it is very difficult for a human scorer 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.

- **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.
- **IntelliMetric is less subject to bias.** IntelliMetric is not affected by the emotional content of an essay response. It is blind to a particularly inflammatory argument or topic. Again, the same cannot be said for human scorers.

What Does IntelliMetric Look at to Score Essays?

One of the most frequently asked questions is, "What does IntelliMetric look at to score essays?" To some extent this is a misguided question. This is akin to asking what do you look at when you make a decision to open a door—certainly the features of the door that are examined are important, but the process for deciding whether or not it is a door is far more important. There is no one "formula" for identifying a door; it is the unique combination of learned features and the organization of those features that lead you to conclude whether or not it is a door.

In a similar vein, what is most important about IntelliMetric is the process it uses to evaluate essay responses. IntelliMetric examines more than 400 features of text, but it is the systemic interaction, or the way in which these features relate to each other together, that produces meaning. A composite picture of the writing is formed from these 400 or so individual elements. Moreover, it is the comparison of this interacting set of features to past learning (from the training phase) that produces meaning.

- **Text Features Examined.** IntelliMetric analyzes more than 400 semantic, syntactic and discourse level features to form a composite sense of meaning as illustrated in the diagram below. These features are then mapped into larger groups we refer to as Latent Semantic Dimensions™ (LSD). These Latent Semantic Dimensions fall into five major categories:
- **Focus and Unity.** Features pointing towards cohesiveness and consistency in purpose and main idea (e.g., unity, single point of view, cohesiveness).
- **Development of Content.** Features of text looking at the content covered, the breadth of content, and the support for concepts advanced (e.g., vocabulary, concepts, support, elaboration, word choice).
- **Organization and Structure.** Features targeted at the logic of discourse, including transitional fluidity and relationships among parts of the response (e.g., introduction and conclusion, coordination and subordination, logical structure, logical transitions, sequence of ideas).
- **Sentence Structure.** Features targeted at sentence complexity and variety (e.g., syntactic variety, sentence complexity, usage, readability, subject-verb agreement).
- **Mechanics and Conventions.** Features examining conformance to conventions of edited American English (e.g., grammar, spelling, capitalization, sentence completeness, punctuation).
- **Latent Semantic Dimensions (LSD).** Based on these more than 400 features, IntelliMetric identifies the underlying semantic dimensions for

a given piece of writing. Fundamentally, IntelliMetric synthesizes broader meanings from many more molecular features. More than 400 features of the text and multiple mathematical models are applied to derive the LSD. The extracted LSDs reflect a multi-dimensional semantic space.

How Does IntelliMetric Use This Information to Score Essays?

To fully understand how IntelliMetric works, it is important to understand how IntelliMetric "thinks", or perhaps more appropriately, how IntelliMetric processes information.

Key Principles. There are five key principles that guide IntelliMetric. They are:

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

Each of these key principles is considered below.

Principle 1: IntelliMetric is modeled on the human brain.

IntelliMetric is designed to emulate the way in which the human brain acquires, stores, accesses and uses information. We refer to this approach as neurosynthetic; i.e., relating to the brain (neuro) and artificially created (synthetic).

The brain is composed of a complex network of neurological pathways. The way in which the brain organizes these neurological pathways and the strength of the connections within these pathways is widely believed to drive thinking and action.

The science and art of creating machines that can think and behave like humans is often referred to as Artificial Intelligence. While there are many definitions of Artificial Intelligence (AI), one interpretation of AI is the ability of machines to think. More specifically, AI, as it is used here, is the ability of a machine to carry out a task or action that requires intelligence and that produces results similar to what might be expected of a human.

IntelliMetric relies on a family of techniques falling under the heading of AI. The specific aspect of intelligence we are interested in is the intelligence ap-

plied by human experts to score and evaluate written text provided by examinees when writing essay question responses. The information contained in the text of an essay is harvested, then organized into a meaningful model by IntelliMetric.

Computer scoring. We often use the term “computer scoring” when referring to automated essay scoring approaches such as IntelliMetric. But the concept of a computer scoring an essay is really a misnomer; the computer does not score an essay per se—it merely reflects what it has been taught by experts and applies acquired information to make a decision in a novel situation.

Principle 2: IntelliMetric is a learning engine

While how we learn is still somewhat of a mystery, we know more about this process than ever before. In developing IntelliMetric, we borrowed liberally from what we know about the human learning process. Although there are many differences of opinion on precisely what constitutes learning, for the purposes of this paper, we view learning as a process of acquiring and organizing information to apply in new situations. This is both a cognitive and action-oriented view of learning.

Learning is central to brain function and plays a large role in the thinking process. Therefore, IntelliMetric was developed to be a learning engine. IntelliMetric learns how to score responses to each question or prompt by reading examples that have been previously scored. It has no built in knowledge base; its wisdom is gained solely from exposure to many examples of essay responses that have been scored by expert scorers. The more than 400 content and structure characteristics of the response described above are associated with the score point assigned.

This learning process is an iterative process. Through an iterative algorithm, IntelliMetric learns how to score accurately. IntelliMetric goes through a repetitive process of applying the information gleaned from each essay example, testing its accuracy at each stage in an effort to improve its scoring accuracy. It gets better and better as it learns more and more from seeing each essay example. It’s almost as if you can hear IntelliMetric saying at some point in the learning process: “Oh, I get it now, ***this*** is what a score of 3 looks like!” and “Oh, I see how this essay is different than an essay with a score of 4.”

Learning over time. Unlike many techniques that have been applied to the scoring of essays, IntelliMetric can learn over time. Much like a baby learns from its mistakes, IntelliMetric is capable of increasing its accuracy over time by seeing its mistakes. This error correction function makes IntelliMetric unique among essay scoring techniques. IntelliMetric relies on a continuous learning model; it continually gets smarter.

Modeling the traditional expert scoring process. IntelliMetric mirrors the scoring process typically used by human scorers. The system learns the underlying rubric and internalizes the characteristics that are important for evaluating responses to the question. Human scorers learn to accurately score student writing through repeated exposure to examples of student writing at each score level. Much like the training of human scorers, IntelliMetric needs to understand the characteristics of each score point. Through repeated to exposure to examples of each score point – a score of 1, 2, 3, etc. – IntelliMetric learns what writing characteristics are important in making an evaluation and how those characteristics are reflected at each score point.

If this process sounds familiar, it should. It is essentially the same process the human brain engages in. The brain acquires information based on experience, organizes this information and applies this knowledge in making decisions. So, too, does IntelliMetric acquire information about how to evaluate essays based on exposure to repeated examples at each of the score points. It then organizes this information into meaningful patterns reflecting the underlying rubric to make a decision about what score to assign to new essays with an unknown score.

Natural language processing. One of the tools used to understand the meaning of the text is called natural language processing (NLP). NLP seeks to understand the meaning of text by parsing the text in known ways according to known rules conforming to the rules of the English language. This is an advanced form of what many of us did in school when diagramming a sentence in English class. Vantage's patented NLP engine is used within IntelliMetric to analyze a response.

CogniSearch. CogniSearch is a technology designed to understand natural language; CogniSearch was developed specifically for use with IntelliMetric and is targeted directly at the accurate understanding of language to support essay scoring. CogniSearch technology uses natural language techniques to analyze student writing. For example, the engine examines sentences in relation to each other to assess coherence , concept threading and focus. Similarly, CogniSearch parses the text to understand parts of speech and how they relate to each other syntactically. This allows IntelliMetric to evaluate the text in relation to expectations for standard written English.

Background knowledge of the English language. Most automated text analysis tools and research seek to evaluate or score text based on a limited, closed corpus of information – typically a few hundred examples of student work written to a specific topic. However, much like any one of us brings a wealth of experience in communications (writing, reading, speaking, and listening) to read a given piece of text, an effective automated text evaluation tool must have a thorough background understanding of the English language.

IntelliMetric possesses a vocabulary consisting of more than 16 million words. More important, this vocabulary is organized as a concept net that retains an understanding of the relationships between and among words. Further, the information on parts of speech (e.g., noun, adjective) and frequency of use are stored as additional information for understanding a piece of writing that IntelliMetric may encounter As an additional enhancement, the concept net includes a thorough understanding of these relationships within and across 37 languages.

The concept net provides a significant leg up in understanding text over other automated essay scoring approaches that rely on simple matrices of words or rely solely on a rules-based parsing of text. For example, IntelliMetric understands that "the computer technician is repairing your computer" is related to "the repair person is fixing the CPU".

Principle 3: IntelliMetric is systemic

IntelliMetric contains many individual pieces of information working in unison to produce a scoring solution that is much more than is represented by any of those individual pieces of information. It is nearly impossible to characterize an automobile in terms of its component parts; they no more add up to a car than do the individual pieces of IntelliMetric add up to an

essay scorer.

Systems theory also tells us that there is more than one way or configuration to arrive at the correct answer. This is important to understanding IntelliMetric. At the risk of oversimplification, different combinations of features taking on different values can all lead to similar scoring decisions. This is in sharp contrast to other attempts at automated essay scoring that rely on purely statistical models. For example, at a gross level, one can achieve a high score with a significant development of well organized content that fails in the areas of mechanics and grammar, or achieve that same score with a somewhat less developed and somewhat less sophisticated organization by excelling in sentence structure.

Nonlinear. Other automated essay scoring systems are based on what statisticians call the General Linear Model. Linear, in this context, means that when looking at two variables, as one quantity increases the other increases a proportional amount. This approach would have us believe that as the values of the text features increase, the score increases in a lock-step fashion in a straight line. This approach is overly simplistic, ignores the complexity of understanding human text and represents a significant departure from a systems approach that recognizes that the understanding of text is both nonlinear and multidimensional.

Principle 4: IntelliMetric is inductive

Inference. You may remember back to grade school that there are two basic types of reasoning: inductive and deductive. Deductive thinking applies a general principle to a specific situation (general to specific); inductive reasoning derives a principle from several example situations (specific to general). Inductive reasoning is based on using several specific instances to form a generalization, whereas deductive reasoning starts with a generalization that is applied to specific instances. They are two different sides of the reasoning coin.

IntelliMetric is largely an inductive process. IntelliMetric makes inferences about how an essay should be evaluated based on its acquired knowledge from specific examples previously evaluated by experts. IntelliMetric models the human scoring process by using information gained from reading the text to make an inference about the score to be assigned. IntelliMetric makes an inference based on several pieces of information in the form of the features of text in the major feature categories described previously. By examining these features of the text, IntelliMetric can make an inference as to what score should be assigned.

Preponderance of evidence. In making inferences, IntelliMetric need not have the complete and absolute answer; it makes use of many sources of information and makes decisions based on the preponderance of evidence. At the core of IntelliMetric are many, many sources of information from which to draw upon to make a judgment about the quality of an essay. Rather than rely on a single source of information, IntelliMetric looks to this variety of sources. The preponderance of evidence is the basis for the decision; all factors need not point to the same evaluation.

Pattern matching. We would simply be overwhelmed with too much information and it would be far too slow if we statically reviewed every piece. We would all like to believe that we carefully process each piece of information available to us and after developing a complete understanding of that information, we take action. On the contrary, it is widely believed that much of

how we think and interpret the world around us is based on pattern matching—a simultaneous interpretation of key pieces of information against a background of historical information to form a reasonable picture.

One area where this process of pattern matching has been studied extensively is the process of human vision. It appears that we create a picture of what we see by filling in the information based on only partial information.

A student's score is a function of a combination of writing features previously identified as important characteristics of student writing. Similarly, IntelliMetric explores the pattern of writing characteristics to provide an evaluation. While any given response is unique, the overall pattern can be matched to the pattern seen for examples at each score point from prior scoring. Much like human judgments, the evaluation of a response emerges from the overall pattern of features seen in the response.

Principle 5: IntelliMetric is multidimensional and nonlinear

Hybrid of techniques. Most attempts at automated essay scoring rely primarily on a single mathematical methodology. Techniques used include linear regression, Bayesian analysis and Latent Semantic Analysis. We recognize the value of these approaches and have incorporated these or related approaches in the development and implementation of IntelliMetric.

Committee Approach. The many approaches used are treated like a committee of judges. IntelliMetric calculates likely solutions (potential scores) from the different mathematical models and sources of information. IntelliMetric then combines this information using proprietary algorithms to obtain the optimal solution, or more simply the solution that is most likely to produce an accurate score. This approach produces the most stable and accurate score possible. In short, rather than relying on a narrow single method and relying on limited information, IntelliMetric draws from several approaches to produce the most accurate results. Since any single judge is more likely to be incorrect, relying on a broader array of information and looking to the optimal solution improves the accuracy and stability of scoring decisions.

How Do We Know IntelliMetric Works?

Over the past seven years, we have conducted more than 150 studies using IntelliMetric. We have compared how well IntelliMetric scored essays to how well human experts scored responses. We looked at how often two experts agreed on what score to assign an essay and compared that to how often IntelliMetric agreed with the experts. In most cases, IntelliMetric was more likely to agree with either expert than two experts were to agree with each other. For example, when examining student responses to an eighth grade writing test, IntelliMetric scores agreed with the experts about 98% of the time; the two experts agreed with each other 96% of the time.

Another way we verified that IntelliMetric works was to compare IntelliMetric scores to the average score across many experts. We looked at how often IntelliMetric agreed with the average expert score and found that the score assigned by IntelliMetric agreed with the average score significantly more often than any individual expert's score agreed with the average score.

Conclusion

IntelliMetric is an advanced artificial intelligence application for the evaluation and scoring of open-ended responses to open-ended essay type questions. Applying a variety of advanced technologies, IntelliMetric scores student responses to open-ended questions as accurately as expert human scorers. Through a complex array of techniques, IntelliMetric can evaluate essay responses, providing virtually instant feedback, dramatically reducing costs, freeing up valuable instructional time and improving student writing.

VANTAGE LEARNING

110 Terry Drive, Newtown PA 18940 / Tel: 215-579-8390 / Fax: 215-579-8391 / www.vantagelearning.com