

How IntelliMetric™ Works

Vantage Learning

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

How IntelliMetric™ Works

“The program in your mind contains a compact description of the world. The objects in the world are elements of that compact description, but they correspond to reality ... because the program is a compact description reflecting training on vast amounts of data.” (Baum 2004, 170)

“...semantics comes from compression...If one compresses enough data into a small representation, the representation captures real semantics, real meaning about the world.” (Baum 2004, 102)

Background and Overview

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. within 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 bear 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 significant 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 1960’s Dr. Ellis Page 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 1960’s, but Dr. Page 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 open-ended questions and provide immediate feedback to students in a matter of seconds.

IntelliMetric™ has been used for a variety of purposes in low and high stakes assessment environments. But arguably the most important application has been in the area of writing instruction. Teachers, schools, state educational agencies, certification programs, and the federal government have been putting more emphasis on improving writing performance through better quality writing instruction. Numerous studies have shown that focusing on writing improvement also brings about gains in other subjects. In short, it is critical that students and professionals are able to write clearly, effectively, and appropriately. In order to do so, it is important for students to have numerous attempts at writing assignments with detailed feedback. In addition, the shorter the time between writing submission and feedback, the more effective and beneficial the feedback is for improving writing.

IntelliMetric. 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. Among the key tools employed in this process are Natural Language Processing, Statistics 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 “learn” the rubric and infer the pooled judgments of the human scorers. The IntelliMetric™ system internalizes the characteristics of the responses associated with each score point and applies this intelligence to score essays with unknown scores.

IntelliMetric™ has begun to have major impact on both classroom instruction and large-scale assessment. With virtually instantaneous electronic scoring, IntelliMetric™ dramatically reduces the cost and time required to evaluate student and professional writing. Moreover, IntelliMetric™ improves the instructional process by offering more frequent and immediate feedback to writers.

IntelliMetric™ shares much in common with the holistic scoring systems commonly employed to score large-scale writing assessments. Typically, 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™ 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™ benefits from the “expert judgments” reflected within the set of papers used to train the engine, not any single scorer’s judgment. Since IntelliMetric™ scoring is a synthesis of many expert opinions it is more reliable (yet may not agree with any single opinion as reflected in a score for a particular paper).

IntelliMetric™ can be used for standardized assessments where a single essay submission is required as well as for various instructional applications where a student can provide multiple submissions of an essay response and receive frequent feedback. IntelliMetric™ Mentor, a complement to the IntelliMetric™ scoring engine, 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 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 this 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 the technology can help. 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 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.

IntelliMetric™: Common Misconceptions

As with any innovation, the novelty of IntelliMetric™ has led to many misconceptions. Before turning to an explanation of how IntelliMetric™ works, let us take a few moments to dispel some of these common misconceptions.

1. ***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.
2. ***IntelliMetric™ cannot “undo” problems caused by poor human scoring.*** Inaccurate human scoring will lead IntelliMetric™ astray; similarly, IntelliMetric™ needs to receive enough papers (100-300) during training to learn how to score correctly. Finally, there must be a sufficient number of papers at each score point on the scale being used to teach the engine (preferably a minimum of 20 at each of the score points). While IntelliMetric™ can mitigate the effects of occasional aberrations in scoring and can do so better than statistically based models, it cannot “make up for” significant errors in the human scoring of training papers.
3. ***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.
4. ***IntelliMetric™ is not magic.*** 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 be fruitful, there is a clear set of processes, well-grounded in theory, that drive IntelliMetric™ that are described below.

5. ***IntelliMetric™ does not focus on surface features.*** On the contrary, IntelliMetric™ examines a complex pattern of more than 400 features that include both relatively straightforward aspects of text such as punctuation and quite sophisticated features such as the expression of concepts. More importantly, as emphasized later in this paper, any single feature is not important; it is the overall emergent pattern that gives rise to meaning.

Why is IntelliMetric™ more accurate than human scorers? IntelliMetric™ is more successful at scoring responses to essay questions than are most human scorers. While IntelliMetric™ still cannot “hold a candle” to the human brain, it does compensate for its limitations in four key ways.

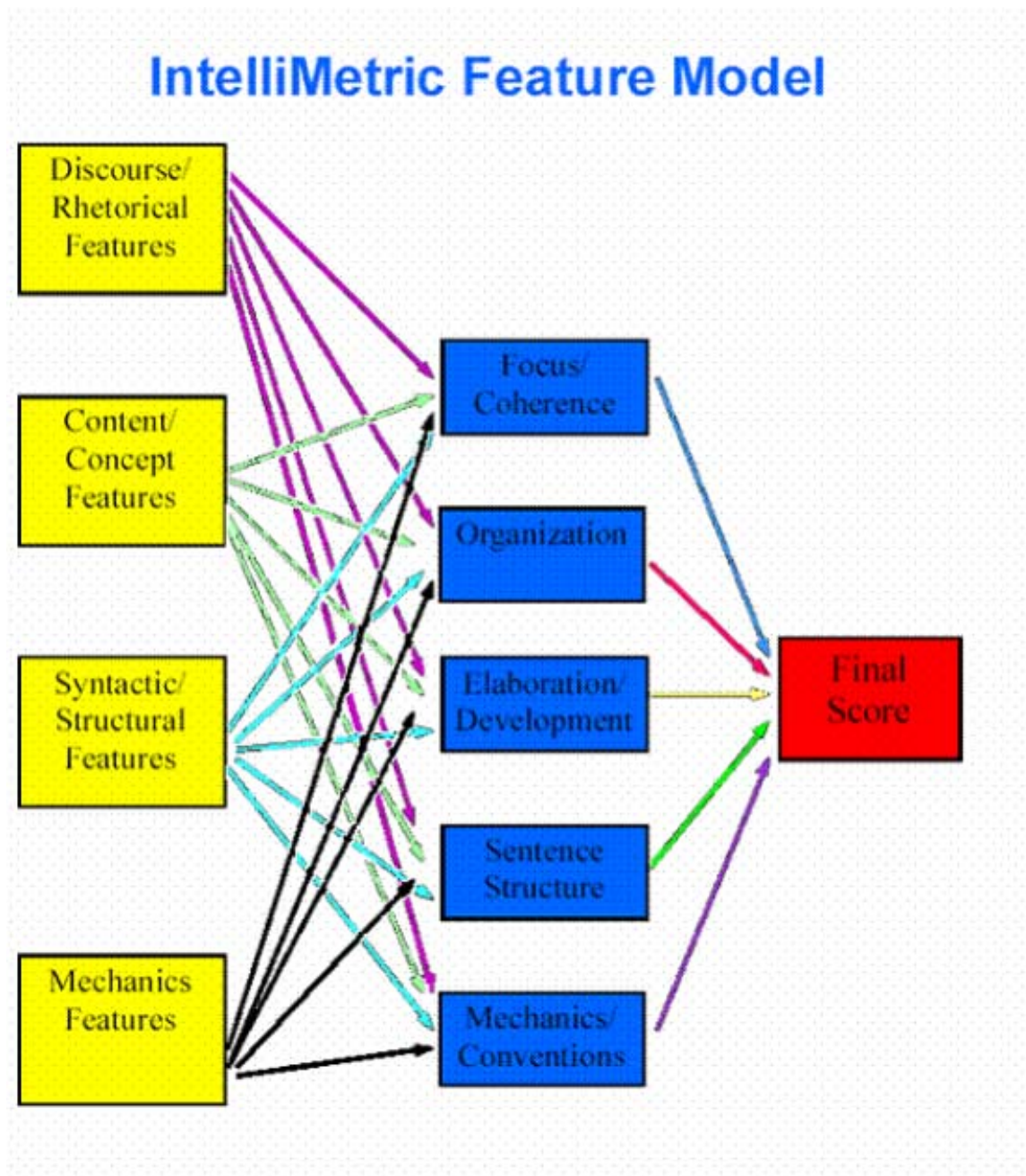
1. **IntelliMetric™ focuses on a narrow domain of understanding.** The human brain must be prepared to solve a vast array of problems in many contexts and domains. This requires the ability to “size up unique situations” and transfer understandings from one domain of knowledge to another. Unlike the human brain, IntelliMetric™ can focus on a very defined domain of understanding defined by a single essay prompt or topic.
2. **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.
3. **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.
4. **IntelliMetric™ is less subject to bias.** IntelliMetric™ is not affected by the emotional content of a given essay response or a particular line of argument that may be offensive or unappealing to a human. 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; not all of the features we associate with “door” need to be present for an individual to recognize it as a door, nor do they need to be present in the exact same “quantity” each time to recognized doors effectively. It is the unique combination of learned features and the remarkable ability of the human brain to see the organizational pattern of those features that lead you to conclude “door” or “not-door”.

In a similar vein, what is most important about IntelliMetric™ is the process it uses to evaluate essay responses. More than 400 features of text are examined by IntelliMetric™, but it is the systemic interaction, or the way in which these features relate to each other, 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 and the prior knowledge base) 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 fall into two major categories: content and structure. Examples of the types of features IntelliMetric™ looks at in each of these categories is provided below.

- **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) Features pointing towards cohesiveness and consistency in purpose and main idea. (e.g., Unity, Single point of view, Cohesiveness) 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)
- **Structure-** Features examining conformance to the conventions of edited American English. (e.g., grammar, spelling, capitalization, sentence completeness, punctuation) Features targeted at sentence complexity and variety. (e.g., syntactic variety, sentence complexity, usage, readability, subject-verb agreement)

Based on these more than 400 features, IntelliMetric™ identifies the underlying semantic structure 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 critical semantic structure of text.

How does IntelliMetric™ use this information to score essays?

There is a long-standing academic curiosity about how the human brain creates meaning and how to model this process. While a review of this literature is well beyond this paper, we make a brief attempt to characterize this nearly two-century tradition in the paragraph below.

Many mark the formal beginning of inquiry into how the mind creates meaning with William James's (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.

Key Principles. In developing IntelliMetric™ we sought to integrate current thinking about the human brain and how the brain processes text to develop meaning. IntelliMetric™ is based on this brain-based model of understanding reflecting several central principles. There are five primary principles that guide IntelliMetric™. 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.
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.

3. **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.
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.
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.

Each of these five 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 artificial intelligence. The specific aspect of intelligence we are interested in here is the intelligence applied 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”, and 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. It is widely believed that we learn to assign meaning—from basic concepts to social patterns of behavior—through our exposures to phenomena and events over time (Schank, 1999; Baum 2004). 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 to new situations. Eric Baum captures this point in stating “...if a compact solution solves a large class of learning problems, it can be expected to be good at solving learning problems in that class which it has not yet encountered.” (Baum 2004, p. 122)

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. Its wisdom is gained primarily from exposure to many examples of essay responses that have been scored by expert scorers. (Although, much like the human brain, this wisdom is complemented by a prior knowledge base of “stored experience”.) 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 example essay. It’s almost as if you can hear IntelliMetric™ saying at some point in the learning process after seeing several examples: “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.”

IntelliMetric™ has no pre-defined set of rules that it uses to score a response; the rubric for scoring emerges from the learning process described above. There is no mechanism for the inclusion of a set of rules in advance; this would be inconsistent with underlying principles of inferential learning.

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 gets smarter.

While IntelliMetric™ has this unique continuous learning ability, this process is often blocked to ensure consistency in scoring over time; IntelliMetric™ is only updated as it is determined that IntelliMetric™ would significantly increase its accuracy based on what it has learned.

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 exposure to examples of each score point- a score of one, two, three, 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 IntelliMetric™ acquires 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 under the name of diagramming a sentence. Vantage’s patented NLP engine (used for the past 20 years in various text processing applications ranging from grammar checking to text search and retrieval) 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 of communication (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 more than 500,000 unique word vocabulary. More importantly, this vocabulary is organized into a 16 million word 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 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. The score is an emergent property of the individual features studied. For example, 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 falls down 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 in a straight line. 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 and ignores the complexity of understanding human text and represents a significant departure from a systems approach which 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; it is inferential rather than rule-governed. IntelliMetric™ makes inferences about how an essay should be evaluated based on its acquired knowledge from specific examples, previously evaluated by experts. Again, 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 above. 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 can make use of many sources of information and make decisions based on the preponderance of evidence. At the core of IntelliMetric™ is an embarrassment of riches—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 use and, after developing a complete understanding of that information, we take action. On the contrary, it is now 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 (forgive the pun) 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.

Greenspan and Shanker (2004) provide an enlightening discussion of the central role of pattern matching in communication. Analyzing infant and child communication, they provide support for the criticality of pattern matching within communication. In short, the developing child learns to interpret a complex array of cues including facial expressions, tone of voice, gestures, postures and later linguistic cues as patterns which lead to the satisfaction of physical and emotional needs.

What is most interesting is the role ignoring information plays in making communication effective. It is not so much the ability to focus on the relevant aspects of a communication, but rather the ability to ignore non-salient information. In fact, the success of interpreting a communication-- whether a letter, an essay or a conversation-- lies in the ability to not only identify the salient information, but ignore information that does not contribute significantly to the overall meaning.

This is among the key features that distinguish IntelliMetric™ from other primarily statistically based models. Unlike purely statistical models that rely on a static set of text features and values consistently applied from response to response, the underlying architecture of IntelliMetric™ is predicated on arriving at judgments that are founded on the preponderance of evidence, *ignoring* information that is not consistent with the pattern observed.

Principle 5: IntelliMetric™ uses multiple judgments based on multiple mathematical models.

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 underlying concepts in the development and implementation of IntelliMetric. But unlike other automated essay scorers, IntelliMetric™ creates several independent judgments, or separate scores.

A panel of experts. The independent judges are treated like a “panel of experts”. In the human essay scoring arena, it is better to have several judgments of the score rather than a single judgment. This is no less true in automated essay scoring. IntelliMetric™ calculates likely solutions (potential scores) from the different mathematical models and sources of information (“electronic experts”). 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 limited information, IntelliMetric™ draws from several approaches to produce the most accurate results. Since any single judge is less reliable than several judges, relying on a broader array of information and looking to the optimal solution improves the accuracy and stability of IntelliMetric™ scoring decisions.

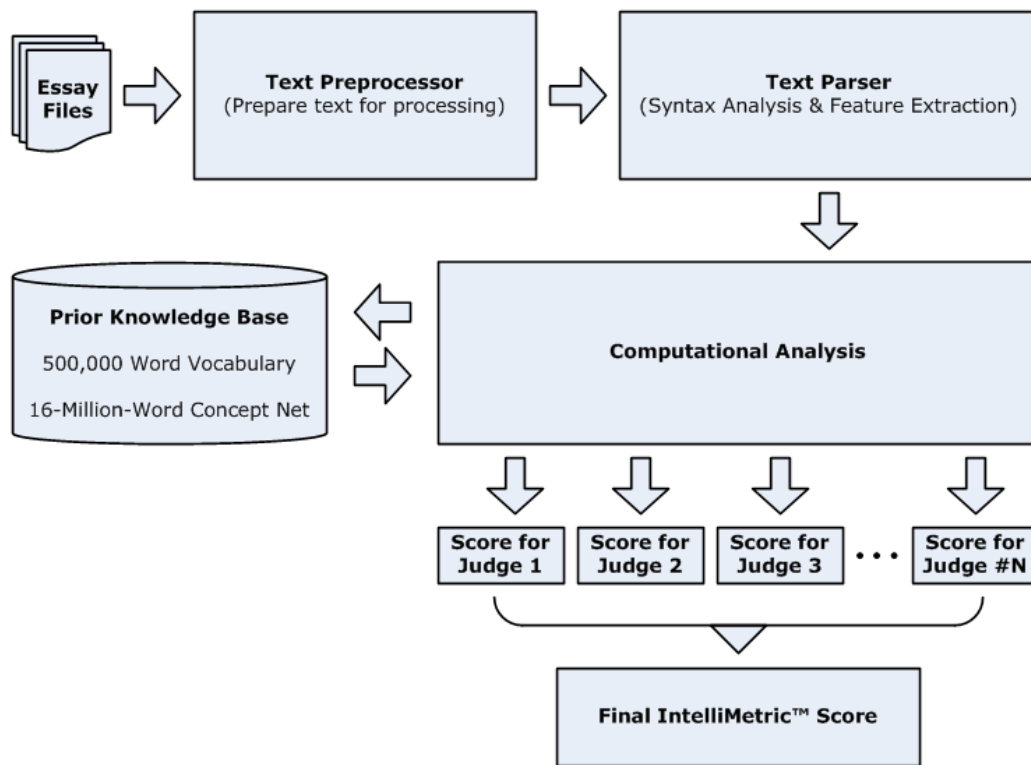
IntelliMetric™ Process

To this point we have examined the theoretical and conceptual basis for IntelliMetric. This section describes the specific process IntelliMetric™ uses to score essays.

Overview of the Process. 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™ evaluates an essay in significantly less than one second; however, to provide a better understanding of how IntelliMetric™ works, this process is broken into steps presented in the following diagram (Figure 1) accompanied by a description of the individual steps.

Figure 1 - IntelliMetric™ Architecture



Step 1: Create essay files. IntelliMetric™ requires that essays be provided in electronic form (ASCII Text). Essay responses can either be transcribed versions of handwritten essays or more commonly essays entered electronically. IntelliMetric™ can accept information as an individual response or as a “batch” of many responses. Increasingly, information is submitted using the Internet as part of a broader educational application, such as MY Access!®.

Step 2: Preprocessing. After the information has been received in electronic form, IntelliMetric™ prepares the information for further analysis. This preprocessing stage makes

sure than all materials are in a form that is readable and understandable by IntelliMetric™. The preprocessor removes extraneous characters and corrects formatting.

Step 3: Analyze text. Once converted to a usable form, the text is then parsed using Vantage’s patented Natural Language Processing engine to understand the syntactic and grammatical structure of the language in which the essay is written. Each sentence is identified with regard to parts of speech, vocabulary, sentence structure, and concept expression. Several patented techniques are used to make sense of the text including morphological analysis, spelling recognition, collocation grammar, and word boundary detection. A 500,000 unique word vocabulary and 16 million word concept net are consulted to form an understanding of the text.

Step 4: Calculate information. After all the information has been extracted from the text, it is translated into numerical form to support computation of the mathematical models. This process relies on a variety of statistical techniques and computational linguistics to create the more than 400 features described earlier.

Step 5: Evaluate text based on virtual judges (Mathematical Models). The information obtained as a result of Step 4 is used as a basis to determine one or mathematical models to make a judgment about the score to be assigned to an essay response. Rather than relying on a single “judge” or mathematical model, IntelliMetric™ employs multiple mathematical judges (“virtual judges”) based on a variety of techniques.

While the number of judges used by IntelliMetric™ varies depending on several factors, they all share certain things in common. At the highest level, each judge seeks to associate the features extracted from the text with the scores assigned in the training set in order to make accurate scoring judgments about essays with unknown scores. They differ with respect to the specific information used to score and more importantly the underlying mathematical model used to make judgments. Several statistical, AI and machine learning methodologies are used to create judges.

In the development stage for a new prompt or topic, this step actually creates the mathematical models or “judges” to be used. After the models have been created, this step would simply apply the mathematical understanding to a novel essay response.

Step 6: Resolve multiple judges’ scores. Step 5 yields several possible judgments. Using a proprietary mathematical model, IntelliMetric™ integrates the information obtained from the judges to yield a single accurate, reliable and stable score.

This is much like human scoring situations where multiple scorers evaluate an essay response and some model must be applied to integrate those diverse opinions.

How do we know IntelliMetric™ works?

Over the past seven years we have conducted more than 200 studies using IntelliMetric™. The studies conducted through about 2001 were summarized in Elliot (2002). We have compared the scores assigned by IntelliMetric™ to the scores assigned by human experts for the same set of essays. 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. We have compared IntelliMetric™ to the experts in studies looking at K-12 students, college admissions candidates, higher education students, and graduate school admissions candidates, to name a few.

In most cases, IntelliMetric™ was more likely to agree with either expert than two experts were to agree with each other. For example, when we looked at 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. These findings vary somewhat from study to study, but all in all, we typically have found that IntelliMetric™ agrees with experts about 95% to 100% of the time—about as often as or more often than experts agree with each other.

Another way we verified that IntelliMetric™ works was to compare the scores assigned by IntelliMetric™ to the average score across many experts. We assumed that the average score of about 8-10 experts was a pretty good estimate of the “real” score for an essay. We looked at how often IntelliMetric™ agreed with the average expert score and found that the scores assigned by IntelliMetric™ agreed with the average scores significantly more often than any individual expert’s score agreed with the average score. In fact, not one of the individual experts did as well as IntelliMetric™ in comparison to this average score.

The third major way we have looked at IntelliMetric™ is in comparison to other ways of measuring writing and language skills. In other words, we asked: Does IntelliMetric™ tend to agree with the evaluations of student skills offered by other measures such as multiple choice tests, independent teacher judgments, etc.? We found that IntelliMetric™ agreed with teachers’ judgments of student writing, student SAT scores, multiple choice writing tests and several other instruments as well if not better than the scores assigned by experts agreed with these measures.

Based on these studies as adapted from Elliot (2002), we know that IntelliMetric™:

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

IntelliMetric™ seems to perform best under the following conditions:

- **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.

References

- Baum, E.B. 2004. *What is Thought?* Cambridge, Massachusetts: MIT Press.
- Blalock Jr, Hubert. 1972. *Social Statistics*. McGraw Hill. p. 405-407
- Joos, M. 1950. Description of Language Design. *Journal of the Acoustic Society of America* 22:701-08.
- Minsky, M. 1986. *Society of Mind*. Cambridge, Massachusetts: MIT Press.
- Osgood, C.E., J. Suci, and P.H. Tannenbaum. 1957. *The Measurement of Meaning*. Urbana, Illinois: University of Illinois Press.
- Schank, R.C. 1999. *Dynamic Memory Revisited*. Cambridge, England: Cambridge University Press.
- Shermis, M. & Burstein, J. 2002. *Automated Essay Scoring. A cross-disciplinary perspective*. New Jersey: Lawrence Erlbaum Associates.