THE IMPACT OF MY ACCESS!™ USE ON STUDENT WRITING PERFORMANCE:

A technology overview and four studies from across the nation

Paul Edelblut, Kenneth Change (張光森)
Vantage Learning / Summit IntelliMetric

9-2 Fl., 76, Section 2, Dun Hua S. Rd., Taipei, Taiwan

ABSTRACT

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 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 1960's, 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 1980's and has been used successfully to score open-ended essay-type assessments since 1998.

MY Access!™ was created by Vantage Learning embedded with IntelliMetric™ scoring engine, MY Access!™ is designed to meet these needs of quality writing instruction for K-12 and higher education with opportunities for frequent writing assignments coupled with immediate feedback.

Introduction to My Access!™

My Access!™ is a web-based instructional writing product that utilizes the IntelliMetric™ scoring engine to provide immediate feedback on submitted essays. MY Access!™ provides holistic scores on a 4 or 6 point scale as well as analytical scores in the areas of Focus and Meaning; Content and Development; Organization; Language, Use and Style; and Mechanics and Conventions. Currently MY Access!™ contains content for grades 4 through higher education including narrative, persuasive, informative, literature, and expository prompts. Additional components of MY Access!™ for adult populations as well as younger populations and new subject areas including science, math, and social studies are also in various stages of development. IntelliMetric™ has been found to be successful in scoring essays written in other languages and MY Access!™ will shortly be incorporating multiple language prompts as well.

An online portfolio is maintained for every student using MY Access!™. All original drafts, scores, revisions, comments from teachers, reflective journal entries, and IntelliMetric™ feedback are accessible at any time. Teachers and administrators are also able to view these portfolios at the individual, class, school, or higher aggregate level.

Key Features of MY Access!™

In addition to the online portfolio of student responses, scores, comments, journals, and teacher comments, MY Access!™ provides additional writing instruction materials and tools.

Students have access to a variety of tools:

- Writer's checklist to help guide the student through the writing process
- Scoring rubrics so the students can self-assess their writing through the process
- My Editor to provide grammatical comments, suggestions, and explanations of rules. This tool is available at multiple levels of difficulty and

language in order to be most effective for the student.
- Word counter to keep track of length of essay
- Word banks to assist in the selection of appropriate words for use in an essay of a particular genre
- Spelling Checker to assist in the proper spelling of words used in the essay
- Venn diagrams and other graphical writing preparation tools to assist in the formulation and organization of ideas to be included in the essay

Teachers have access to a variety of reports to view the students' writing and feedback in almost any manner. In addition, the teacher has ultimate control over the tools available to the students while writing essays. For example, if it is important that the students do not receive any help with spelling, the spell checker can be turned off for any particular assignment.

Lastly, administrators have access to customized reporting to get just the information they need. Frequency distributions, historical summaries, and roster reports cover just a few of the options.

How MY Access!™ works.

MY Access!™ utilizes IntelliMetric™, Vantage's proprietary automated essay scoring system, to provide scores and feedback on essays. Students are able to revise essays based on the feedback received and submit for a new evaluation of the essay. This process of writing, receiving feedback, continuing to revise, and receiving more feedback helps students improve their writing skills. The value of MY Access!™ would be limited without the implementation of immediate IntelliMetric™ feedback.

Prior to reviewing and interpreting research on the effects of the use of MY Access! ™, it is necessary to have an understanding of how IntelliMetric™ is trained to score essays and how it applies that knowledge to score new essays. Therefore, we turn to a review of how IntelliMetric™ works.

Understanding IntelliMetric™ Automated essay scoring

> *"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)*

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 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™ can not 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™ can not 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™ can not "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 can not "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 can not "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 can not be said for human scorers.

<u>What does IntelliMetric™ look at to score essays?</u>

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.

<u>Text Features Examined.</u> 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.

<u>How does IntelliMetric™ use this information to score essays?</u>

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 this area of inquiry with William James' (1890) fundamental work in association. Inquiry into understanding continued through the early part of the twentieth century with the behavioral movement and slipped into 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) captures the perspective embodied by IntelliMetric™ in his "Society of Mind" view of the brain; here, understanding is seen as the result of thousands of millions 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 produced an integrated view of meaning best reflected in the quotes presented at the beginning of this section.

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", then organized into a meaningful model by IntelliMetric.

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.

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.

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.

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.

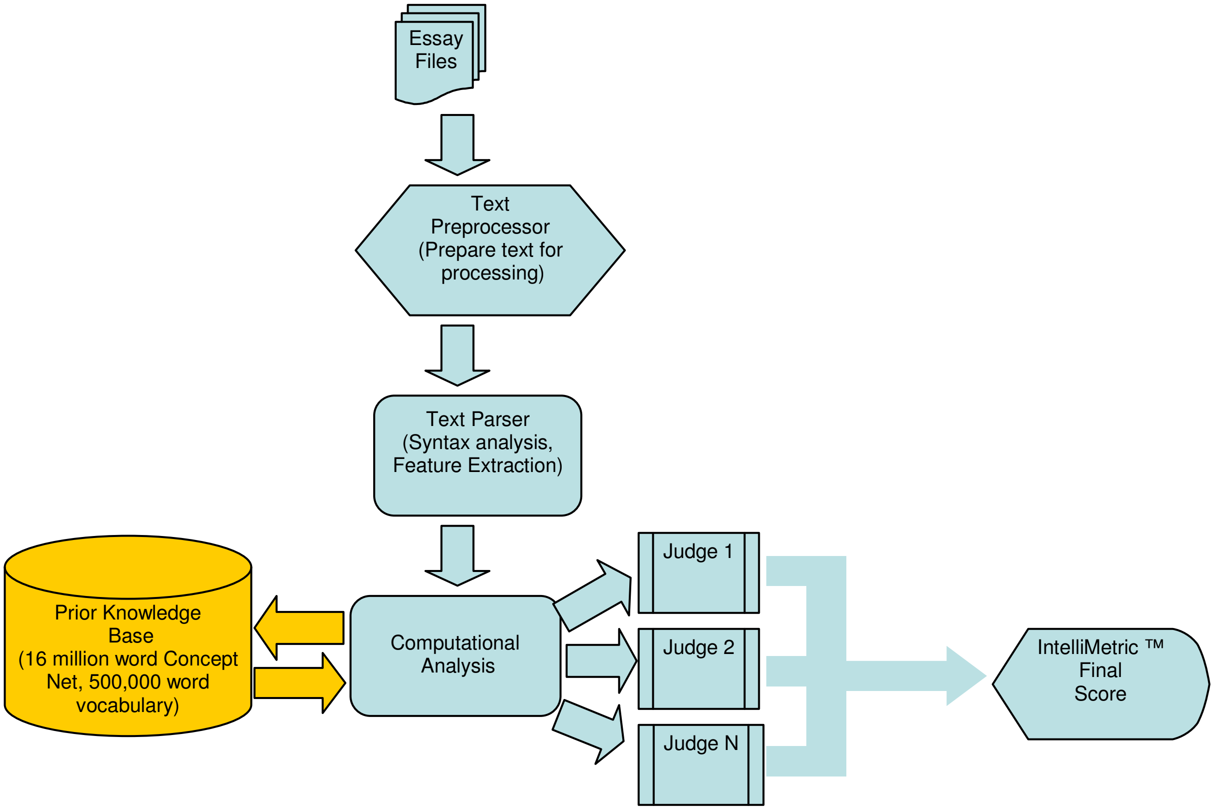
IntelliMetric™ Process

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


Essay Files
Text Preprocessor (Prepare text for processing)
Text Parser (Syntax analysis, Feature Extraction)
Prior Knowledge Base (16 million word Concept Net, 500,000 word vocabulary)
Computational Analysis
Judge 1
Judge 2
Judge N
IntelliMetric ™ Final Score

Step 1: Create essay files. .

Step 2: Pre processing. .

Step 3: Analyze text.

Step 4: Calculate information.

Step 5: Evaluate text based on virtual judges (Mathematical Models)

Step 6: Resolve multiple judges' scores.

How do we know IntelliMetric™ works?

Over the past 7 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 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:

- ***Larger number of training papers:*** 300+ (although models have been constructed with as few as 50 training papers).
- ***Sufficient papers defining the tails of the distribution:*** For example on a one to six scale it is helpful to have at least 15 papers defining the "1" point and the "6" point. (Although, models have been constructed with few or no papers at the extremes).
- ***Larger number of expert scorers used as a basis for training:*** Two or more scorers for the training set seem to yield better results than 1 scorer.
- ***Six point or greater scales:*** The variability offered by six as opposed to three or four point scales appears to improve IntelliMetric™ performance.
- ***Quality expert scoring used as a basis for training:*** While IntelliMetric™ is very good at eliminating "noise" in the data, ultimately, the engine depends on receiving accurate training information.

Under these conditions, IntelliMetric™ will typically outperform human scorers.

In addition to this IntelliMetric research across applications, we have conducted many studies regarding the use of MY Access!™, which incorporates IntelliMetric™ scoring, feedback, and tools. The next section of this paper presents four studies regarding the impact of MY Access!™ use on writing performance.

Study 1: A preliminary study of MY Access!™ impact on writing performance: Whittier Union High School District, California

Study Design

This preliminary study provides an initial view of the efficacy of MY Access!™ for use in improving student writing as part of developmental K-12 coursework. A single class of 25 students was selected to participate in the study.

Participants. Twenty five grade nine students participated in the study. These students were part of a Summer School program offered in an urban school district for at risk students. At risk students were those who performed poorly academically, exhibited poor writing and/or had poor standardized test results in the language arts area. These students were seen as at risk for continued poor academic success and potential failure on statewide writing assessments.

Design. The study was conducted from July 1, 2001 to August 18, 2001. All twenty five students were provided with passwords and access to MY Access!™ and were informed that MY Access!™ would be used as a core component of the Summer School program with the goal of improving their writing skills. Students were provided with specific writing assignments on an approximately weekly basis and were also encouraged to do additional writing at their own discretion. Approximately 4 writing prompts were assigned over the course of the study. Students typically wrote between 2 and 5 revisions of each of the assigned prompts.

Two content parallel measures of direct student writing (two separate writing prompts scored on the same rubric) were used. Student performance on their first response was compared to their performance on their last submission as a measure of writing improvement. Writing was measured on a 4 point scale using an established rubric that included: Organization, Development, Focus, Sentence Structure and Mechanics. Student performance was compared overall as well as on five dimensions of writing (see below).

Feedback/Scoring. Students were provided feedback on a four-point scale based on the MY Access!™ rubric. For each response submitted, students received an overall score, as well as a score in focus, organization, development, sentence structure and mechanics. Students also had access to sample high quality responses and additional instructional information through the My Access! application.

Results

The average score obtained initially by students and the final score obtained as described above were compared as a measure of writing improvement.

.

Discussion

In a span of just over 6 weeks, students showed significant gains in writing performance. Students gained, on average, between one half and one full point on the four point scale. Given the narrow range of the scale (1-4) this represents substantial growth. One point represents 25% improvement in performance. Perhaps more importantly, students moved from about the “2” point on the scale to the “3” point on the scale; in typical statewide student assessments this often represents the difference between failing and passing.

At risk students in Summer School programs such as this represent a significant instructional challenge. Minimal growth is generally seen in such programs. The magnitude of the growth seen in this preliminary study suggests that MY Access!™ is a valuable tool for writing development instruction.

Follow up.

Based on the success of the 2001 Pilot Program, Whittier expanded the use of MY Access!™ to all ninth- and tenth-graders throughout the district. Standards-based teaching coupled with MY Access!™ writing development instruction had sizeable impact on increasing the API scores for the district’s high schools.

Significant gains in writing performance were seen between 1999 and 2003 (see Tables 1 and 2 below). With the assistance of MY Access!™, Whittier’s 2003 API scores increased to a range of 610 to 687 from former 1999 baseline levels of 480 to 601. The portion of the gain in API scores from 2002 to 2003 accounted for anywhere from 40% to 87% of the total gains since 1999. The two biggest improvements occurred in Pioneer High School (PHS) and Whittier High School (WHS), the two schools in the district with the highest minority and economically disadvantaged populations.

The Academic Performance Index (API) is the cornerstone of California's Public Schools Accountability Act of 1999 (PSAA). The purpose of the API is to measure the academic performance and growth of schools. It is a numeric index (or scale) that ranges from a low of 200 to a high of 1000. A school's score on the API is an indicator of a school's performance level. The statewide API performance target for all schools is 800. A school's growth is measured by how well it is moving toward or past that goal. A school's base year API is subtracted from its growth API to determine how much the school improved in a year. The performance indicators contributing to the API include:

- Standardized Testing and Reporting (STAR) program
  - Norm-referenced test (NRT) - all content areas
    2002 API Base: Stanford Achievement Test, Ninth Edition (Stanford 9)
    2003 API Growth: linked California Achievement Test, 6th Edition Survey (CAT/6)
  - California Standards Tests (CSTs) - English-language arts, mathematics, history-social science (science to be added later)
- California High School Exit Examination (CAHSEE)

While it can not be said with certainty that gains in performance were due to MY Access!™ use, it is clear that schools using MY Access!™ as part of a writing program show significant gains in performance over time.

Table 1

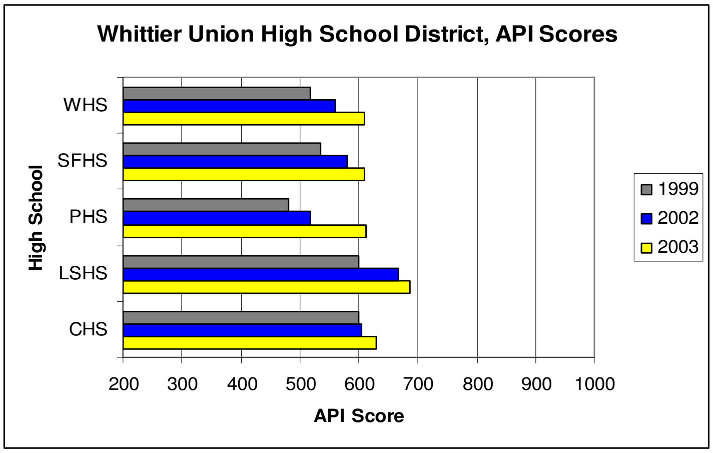
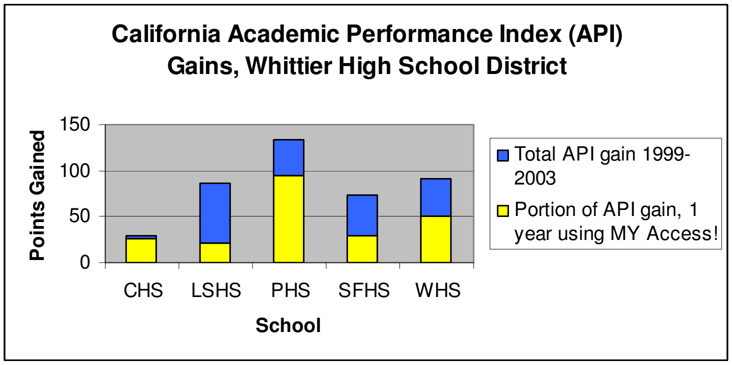


Table 2



Study 2: Birmingham High School, Los Angeles Unified School District, California

Birmingham High School is one of 51 schools in District C of the Los Angeles Unified School District in California. Although the state agreed to delay the requirement of passing the California State High School Exit Examinations (CAHSEE) until 2006, Birmingham administrators and educators knew they needed to take immediate action in order to raise their students’ achievement levels to acceptable levels.

Description of the population/sample. Approximately 75% of the students enrolled in Birmingham High School are economically disadvantaged, and 65% are also English Language Learners.

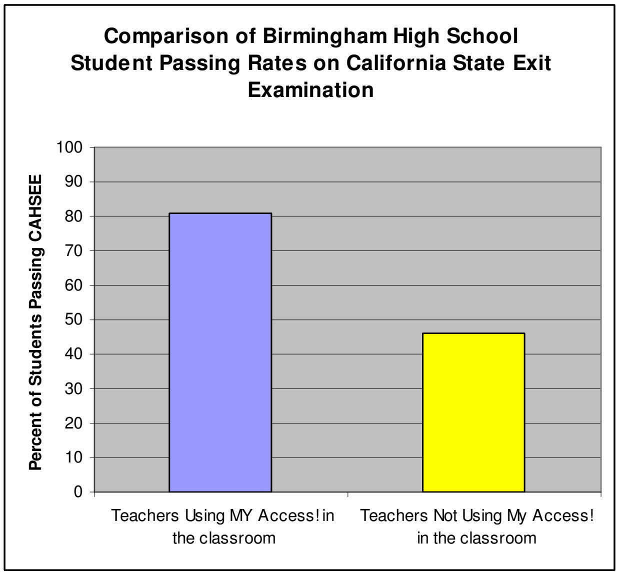
Approximately 812 tenth-grade English Language Learners (ELL) were selected to participate in a study of the impact of using the MY Access!™ writing application. Approximately half of the classes representing about 496 students used the MY Access!™ application while the remaining 306 students were in non participating classes. Participants were determined on a voluntary basis (not random assignment).

Procedures. Those classes that elected to use MY Access!™ did so between approximately October of 2002 and April of 2003 (prior to the CAHSEE test administration). The remaining non-using classes were provided with the existing English Language Arts instruction during the same time period. Students used the MY Access!™ application between 2 and 5 times per week as part of their regular English Language Arts instructional time.

All students participating in the study took the required California High School Exit examination (CAHSEE) in April of 2003. The CAHSEE performance of those students using MY Access!™ was compared to those students who did not use MY Access!™.

Results. Eighty one percent of Birmingham students who used MY Access!™
(N passing = 405; Total N = 496) passed the California High School Exit Examination, while only 46% of the students who did not use MY Access!™ (N passing = 142; Total N = 306) passed the exam. A chi square test comparing the two passing rates was significant ($\chi^2 = 108.42$, $p< 0.001$).

Table 3



Discussion. MY Access!™ appeared to have a sizeable impact on the writing skills of the students using the program in their classrooms. The CAHSEE passing rates of individual classes using MY Access!™ ranged from a low of 60% (in a classroom with only 10 students) to 100% (in a classroom with 49 students). The majority of pass rates were in the mid-70s to mid-80s range. These results are illuminating, especially given the number of participating students.

With about 800 students studied, the effect does not appear idiosyncratic. All the same, this study did not rely on random assignment and it is possible that self selection in to the participating/non-participating groups could account for the variance in performance of the two groups.

Study 3 Red Clay Consolidated School District, Delaware

The Red Clay Consolidated School District in New Castle County, Delaware consists of 28 schools in a predominantly suburban and rural setting. Red Clay Consolidated School District piloted the use of MY Access!™ in selected elementary and middle schools during the 2002-2003 school year.

After only one year of using MY Access!™ at selected schools, those schools showed significantly higher levels of writing proficiency as measured by the statewide writing assessment examination, the DSTP Writing Performance Levels.

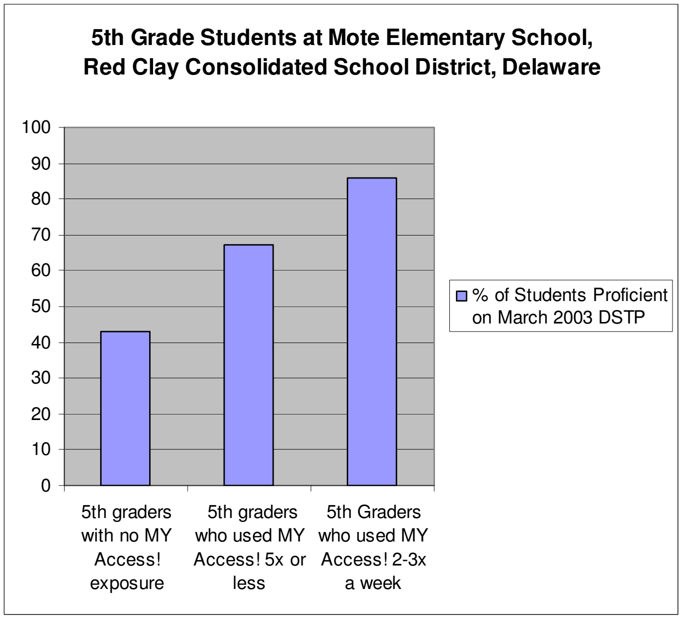
Description of the sample. Approximately 195 fifth-grade students were selected to participate in a study of the impact of using the MY Access!™ writing application. Approximately half of the classes representing about 100 students used the MY Access!™ application while the remaining 95 students were in non participating classes. Participants were determined on a voluntary basis (not random assignment).

Procedures. Those classes that elected to use My Access! did so between approximately October of 2002 and March of 2003. The DSTP was administered in March 2003 and was used as a criterion measure for comparing MY Access!™ users and non users. Non-using classes were provided with the existing English Language Arts instruction during the same time period. Students used the MY Access!™ application between 2 and 5 times per week as part of their regular English Language Arts instruction.

All students participating in the study took the required DSTP writing examination in March of 2003. The number and percentage of students achieving the highest two levels on the DSTP within the group using MY Access!™ were compared to the number and percentage of those achieving those levels within the groups which did not use MY Access!™. Students were also asked to indicate their frequency of use of MY Access!™ to allow further comparisons. Those using MY Access!™ 2-3 times per week were compared to those who used the application 4 or 5 times per week.

Results. Fifth-grade students who used MY Access!™ had sizeable academic gains on the March 2003 DSTP Writing Assessment. 86% of fifth-grade students who used MY Access!™ two or three times a week scored at Performance Level (PL) 3 or 4 on the DSTP; 67% of fifth-grade students who used MY Access!™ a total of four or five times per week scored at PL 3 or 4. Only 43% of fifth-grade students who had no exposure to MY Access!™ scored at PL 3 or 4. Chi Square comparisons showed that these relationships were significant (χ^2= 24.32, p < 0.001).

Table 4



Discussion. MY Access!™ appeared to have a sizeable impact on the writing skills of the students using the program in their classrooms. The DSTP passing rates were significantly higher for those students who used MY Access!™. It is interesting to note however, that those using the application 2-3 times per week seemed to perform better than those who made use of the application more often.

Again, while these results clearly point to strong effects for MY Access!™ use on later writing performance, the lack of random assignment suggests that caution should be used in interpreting the results; it is possible that self selection in to the participating/non-participating groups could account for the variance in performance of the two groups.

Study 4: Parkland High School, Pennsylvania

Parkland High School is located outside of Allentown, PA and after administering the Spring 2002 PSSA (Required State) Writing Test, administrators and educators found an unacceptable 22% of all students scored Below Proficient. In response to this perceived problem, MY Access!™ was adopted as one way to address the challenge for the 2002-2003 11th grade class.

Study Design. In the Fall of 2002, the approximately 709 11th grade students were recruited to participate in a pilot program and research study of MY Access!™. Students used MY Access!™ between September 2002 and March 2003. Students took the PSSSA writing assessment in March 2003. The 2003 PSSA results for (MY Access!™ use) were compared to performance for that class on the 9th grade PSSA writing test in 2000 (with no MY Access!™ use).

Results

The MY Access!™ treatment had sizeable impact in reducing the number of students who were at risk of not meeting Proficient standards (see Table 5 below). Using MY Access!™, 91% of 709 students tested on the PSSA in 2003 attained scores of Proficient and Advanced, compared to only 76% of 711 students tested in 2000 that did not use MY Access!. Using MY Access!™ less than 10% of the class in 2003 is rated Below Proficient on their PSSA scores, compared to 25% of the class in 2000 that did not use MY Access!™.

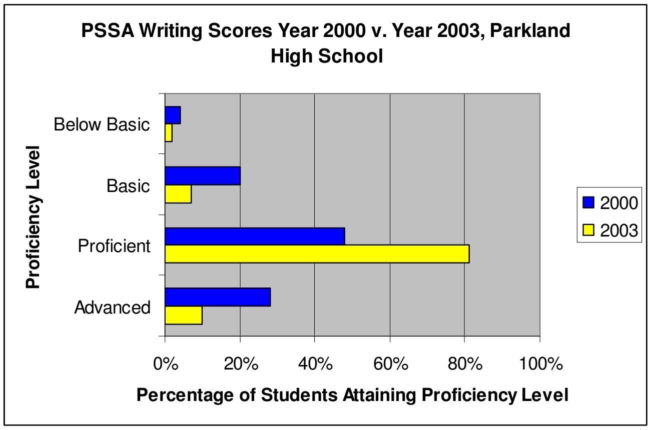
MY Access!™ had a sizeable impact on raising PSSA test scores from Basic to Proficient, or from Below Basic to either Basic or Proficient. Fifteen percent more students demonstrated Proficient and Advanced writing skills in 2003 than in 2000. In other words, 100 more students tested on the PSSA in 2003 after using MY Access!™ attained scores of Proficient and Advanced than in 2000 when MY Access!™ was not used.

Discussion

As with the other three studies cited, MY Access!™ appeared to have a sizeable impact on the writing skills of the students using MY Access!™. PSSA proficiency rates were up significantly when compared to the year 2000 baseline.

While these results clearly point to strong effects for MY Access!™ use on later writing performance, the lack of control between the baseline year of 2000 and the year 2003 suggest that the effects could be attributable to other factors. Moreover, while the state makes strong efforts to achieve comparability in evaluation of essays from 9th to 11th grade, these results may not be directly comparable.

Table 5



In addition to these studies investigating the effects of MY Access!™ use on writing performance, we are also interested in the student perceptions of using MY Access!™ in the classroom. The following describes one such study.

This particular study provides a view of students' perceptions of using MY Access!™ based on a questionnaire completed during Spring 2003. Students used the MY Access!™ application for approximately 6 months between October 2002 and March 2003.

Participants. Ninety-four eighth grade students were asked to participate in the study. These students were enrolled in the Hilton School District in New York State.

Design. All 94 students were provided with passwords and access to MY Access!™ and were informed that MY Access!™ would be used as a core component of the writing program with the goal of improving their writing skills. Students were provided with specific writing assignments approximately on a weekly basis and were also encouraged to do additional writing at their own discretion. Students typically completed an initial draft of a response to each prompt and did between two and five revisions of each essay.

At the end of March, students were asked to complete a multi-question survey of their attitudes and opinions about MY Access!™. The frequency and percentage of students selecting each response to each question were calculated (see below).

Results

Overall Perceptions. Overall, students felt very positive about using the MY Access!™ product. Almost all (87%) of the students recommended that the English Language Arts teachers use the program again. More than four-fifths (81%) of the students indicated that MY Access!™ helped them prepare for the required state English examination.

Perceptions of Writing Improvement. More than four-fifths (81%) of the students felt that MY Access!™ helped them improve their writing. Similarly, more than four-fifths (82%) of the students indicated that they used the feedback provided to improve their writing.

Perceptions of Scoring Accuracy. More than four-fifths (83%) of the students thought that the MY Access!™ scoring was fair and accurate. Nearly three-quarters (73%) of the students found the scoring feedback to be good, very good or excellent and more than four-fifths of the students said they used the feedback to improve their writing scores.

Discussion

Clearly, students found the MY Access!™ program to be helpful. More than four-fifths of the students felt the program and the feedback it provided helped them improve their writing. Nearly all recommended continued use of the product and saw the product as helpful in preparing for the statewide high stakes examination. The scoring was generally seen as fair and accurate.

Students' perception of curriculum materials is very important in the evaluation of any instructional strategy. While student perceptions alone are insufficient for evaluation, these results combined with previous studies demonstrating significant writing skill improvement suggest that MY Access!™ is a beneficial component of a school's writing program.

Summary

The research presented indicates that students who use MY Access!™ tend to show greater improvement in writing and language arts skills than those students who do not use MY Access!™. More specifically, when students use MY Access!™ several times a week they show dramatic gains in writing skills as measured by high stakes statewide examinations and other standardized writing measures. In addition, surveys of student perceptions of MY Access!™ have been overwhelmingly positive with one example presented in this paper.

In summary, these studies demonstrate:

1. ***Students show significant gains in writing skills (as measured by parallel forms of the same instrument from pre to post test) over time.***
2. ***Students show significant gains on statewide high stakes examinations after approximately 6 months of using My Access!.***

All the same, these results need to be considered preliminary. There are several limitations that caution against drawing these conclusions too strongly. While there is no reason to assume a bias in assignment of students to MY Access!™ Use or No MY Access!™ Use treatment groups, students were not randomly assigned to conditions.

It is generally unknown what treatments non MY Access!™ users were provided. This however, is more likely to mitigate the effects seen here. These "non-using students" were not subject to ***no*** treatment; rather, they presumably received an alternative form of instruction during this period (In simple terms, they did not "just sit there" while others were using MY Access!™.)

Future research is being targeted at larger studies under more controlled conditions with random assignment. Also, research will be conducted on the effectiveness of MY Access!™ for writing instruction in areas outside the K-12 and higher education areas.

References

Baum, Eric B. (2004) What is Thought? MIT Press: Cambridge, Massachusetts.

Elliot, S. (2002) From Here to Validity in, Shermis, M. and Burstein, J. Automated Essay Scoring. New Jersey: Lawrence Erlbaum Associates.

Joos, M. (1950) Description of Language Design. Journal of the Acoustic Society of America., 22, 701-708.

Minsky, Marvin (1986) Society of Mind. MIT Press. Cambridge, Massachusetts.

Osgood, C.E., Suci, J., and Tannenbaum, P.H.. (1957) The Measurement of Meaning Urbana, Illinois: University of Illinois Press.

Schank, Roger C. (1999) Dynamic Memory Revisited. Cambridge, England: Cambridge University Press.