A Multi-Method, Multi-Trait Validity Study of Direct Writing Assessment Using Automated Essay Scoring

Bo Wang  
The College Board

Cathy Mikulas  
Vantage Learning

A Multi-Method, Multi-Trait Validity Study of Direct Writing Assessment Using Automated Essay Scoring

This paper presents the results of a three-year study of a direct writing assessment program incorporating automated essay scoring technology used for college placement. The study, conducted at more than 100 colleges and universities throughout the United States, examines the validity of the instrument using a multi-trait multi-method approach. Several hundred thousand students took the direct writing assessment and one or more other measures as a basis to examine the construct validity of the writing component.

The writing assessment was found to correlate more highly with other measures of the writing construct, less so with other measures of language arts and least of all with measures in the mathematics domain. The results provide additional support for a growing body of evidence supporting direct writing assessment using automated essay scoring.

With the growth in interest in writing assessment at the K-12 and post-secondary levels and the rapid increase in the use of automated essay scoring technology, a continuing analysis of validity is critical.

This study examines the validity of the test scores for WritePlacer®, a direct assessment of writing included as one component of The College Board’s ACCUPLACER® college placement testing program. A multi-method, multi-trait validation approach is used to investigate the extent to which WritePlacer® scores relate to other measures of writing and other academic skill areas. The findings confirm that WritePlacer® is significantly associated with other measures of writing, moderately associated with other measures of reading and less so with measures of mathematics. These results are seen as contributing to the overall validity evidence in support of the use of WritePlacer® scores.

About ACCUPLACER® and WritePlacer®

ACCUPLACER® is a computerized placement testing system used to place students in appropriate college courses. ACCUPLACER® provides several assessment options in the areas of Reading, Writing, and Mathematics. Several multiple choice computer adaptive measures are used along with a direct assessment of student writing. ACCUPLACER® is used by nearly 1,000 higher education institutions in the U.S., Canada, Europe and Asia.

One component of the ACCUPLACER® program is WritePlacer®. WritePlacer® is a direct assessment of writing used to place entering college students into appropriate English courses. Students are presented with a prompt, identifying a topic and a charge for writing and are asked to provide a writing sample of approximately 300 words in response. Students are provided with an immediate score using IntelliMetric™ automated essay scoring technology. (In approximately 5% of the cases an essay is flagged, a pending score is received and expert scorers determine and provide a final score within 24 hours.)

WritePlacer® is scored on a 6 point rubric reflecting 5 dimensions of writing as determined by College faculty. The five dimensions are: Focus, Organization, Development, Sentence Structure and Conventions. Alternatively, WritePlacer® is offered on a 4 point scale, reflecting the same domains of writing in Texas. WritePlacer® was initially scored by trained expert scorers; this information was then used to train the IntelliMetric™ engine and is used to score student responses.
About IntelliMetric™

Automated essay scoring technology refers to the evaluation or scoring of open ended, constructed response assessments by electronic means. There are several methodologies currently being employed; they all trace their roots back to Ellis Paige’s work in the 1960s with Project Essay Grade (Burstein and Shermis, 2002). Models include the e-rater developed by ETS and the Intelligent Essay Assessor (Burstein and Shermis, 2003). Currently, the most widespread used methodology is IntelliMetric™ which has been in use operationally since about 1998. IntelliMetric was employed in the current study. A brief overview of the IntelliMetric technology is provided below; a more complete technical discussion is provided in Appendix A of this paper.

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™ 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).

The IntelliMetric™ tool provides feedback on overall performance, diagnostic feedback on several rhetorical and analytical dimensions of writing (e.g., conventions, organization), and provides detailed diagnostic sentence-by-sentence feedback on grammar, usage, spelling and conventions.

More than 400 semantic, syntactic and discourse level features of text are examined by IntelliMetric™; the systemic interaction of these features is used to form a meaningful composite picture of the writing.

Over the past 8 years more than 200 studies have been undertaken using IntelliMetric™. Researchers have compared the scores assigned by IntelliMetric™ to the scores assigned by human experts for the
same set of essays. They have 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. Researchers 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 researchers 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, studies 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.

Based on these studies as adapted from Elliot (2002), it can be concluded 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.

See Appendix A for a more detailed discussion of IntelliMetric.

**Prior Studies of WritePlacer™ IntelliMetric™ Scoring**

Several studies exploring the use of IntelliMetric automated scoring technology with WritePlacer® have been conducted since 2000. The results suggest that IntelliMetric scoring of WritePlacer® essay responses is comparable to the scores seen with expert human scoring. A summary of some of the key findings of these studies is presented below.

The IntelliMetric scoring of the WritePlacer® program was first the subject of a study conducted in 2000 (Vantage Learning, 2000). In this study, IntelliMetric™ agreed with the expert graders within one point 100% of the time, exactly with scorer one 76% of the time and exactly with scorer two 80% of the time. These figures compare favorably with the expert scorer to expert scorer agreement of 100% (adjacent) and 78% exact.

In 2000, 98 students at a large Mid Atlantic College participated in a validity study of WritePlacer® (Vantage Learning, 2002a). The essays were scored by college faculty, external experts and automated scoring technology. These scores were correlated with each other and with student SAT results.

The results showed that the automated essay scoring matched expert scorers 92% of the time, while college faculty matched the automated essays scores 43% of the time. External expert scorers matched college faculty scoring 54% of the time. A Pearson r correlation of .92 was observed between automated essay scores and external expert scoring, and .64 between automated essay scores and with college faculty. College Faculty scores correlated with expert scores .60 (Pearson r; p<.01)

IntelliMetric™ was found to correlate more highly with external measures of Verbal skills than did either expert scorers or college faculty. IntelliMetric™ was more highly correlated with SAT results than with expert or faculty scoring. The Pearson R correlation between IntelliMetric scores and student SAT scores was .51. SAT scores showed a .47 correlation with college faculty scores and a .42 correlation with expert scores .42 (Pearson R; p<.01).
A 2000 WritePlacer® validity study conducted at a large Southwestern University found a .72 Pearson r correlation between faculty assigned scores and IntelliMetric scores (Vantage Learning, 2001). The results also show that 91% of the essays were scored the same or within one point agreement.

In 2001, a study comparing computer essay scoring of WritePlacer® and human expert essay scoring of WritePlacer®, was conducted at two large New England community colleges (Vantage Learning, 2001). The correlations among two faculty scorers and IntelliMetric scoring found that while faculty scores correlated with each other .66, IntelliMetric correlated with faculty #1 .67 and with faculty #2 .69. Finally, IntelliMetric scores showed a .77 correlation with the average score of the faculty members.

In 2002, several prompts targeted for use with ESL populations were added to the WritePlacer® program (Vantage Learning, 2002b). Expert human scoring was compared to IntelliMetric scoring for both the overall scores and five dimensional scores. Overall, expert scorers and IntelliMetric agreed within one point 94% to 98% of the time with correlations ranging from .78 to .84.

A 2002 study of IntelliMetric™ scoring accuracy for the WritePlacer® program found significant agreement between expert human scorers and IntelliMetric™ (Vantage Learning, 2002a). When examining two WritePlacer® program prompts, IntelliMetric™ scores and those produced by expert human scorers were found to agree within 1 point 100% of the time and agree exactly between 72% and 74% of the time. The correlation between IntelliMetric™ and expert graders was .71 for both prompts examined.

A 2004 replication of the WritePlacer® IntelliMetric™ scoring accuracy supported earlier findings. The use of the most recent release of IntelliMetric™ (9.2) seemed to further improve results (Vantage Learning, 2004). A single WritePlacer® prompt was evaluated and served as the basis for a study reexamining scoring accuracy (Vantage Learning, 2004). The results suggest that the mean score produced by expert scorers and IntelliMetric™ were comparable with no significant difference found (t=-.382 p>.05). The human expert and IntelliMetric scorers agreed exactly 68% of the time and within 1 point 100% of the time. The relationship between the human expert scores and IntelliMetric™ scores was quite high (Pearson r= .81).

The number of training papers and the distribution of training papers is often a concern in developing IntelliMetric™ scoring models. A follow up study examined the impact of the training set size on the scoring accuracy (Vantage Learning, 2004). Three different training set conditions were examined: 300, 200 and 100 training papers selected at random. In all three training set conditions, the average agreement rates were at or above 95%. And the correlation between IntelliMetric™ and human scores was .83 or above. Larger training set sizes tended to produce the most accurate results. The 300 paper and 200 paper training sets produced agreement rates of 97% and 98% respectively, with average correlations of .89 and .87 respectively. Some “fall off” was seen with 100 paper training sets; an average 95% agreement rate and an average correlation of .83 was found.

While the research exploring the use of IntelliMetric™ scoring for WritePlacer® conducted to date certainly supports the validity of this automated scoring procedure, establishing the validity of the WritePlacer® IntelliMetric scoring warrants additional study.

Validity

Traditionally, validity has been divided into three primary areas: Content Validity, Criterion-Related Validity and Construct Validity. Content Validity refers to the extent to which a measure relates to the
content domain it is claimed to measure. Criterion-related validity relates to the extent to which a measure correlates with other criteria it is expected to relate to or predict. Construct validity reflects how well a given measure relates to the construct it is claimed to measure.

While useful, this division masks a central tenet of validity: Validity is a process of accumulating various sources of evidence to support the use of test scores; validity is not a study or even several studies; it is an ongoing process of evidence collection. An increasingly prevailing view is that validity is a unified concept with evidence collection targeted at demonstrating a relationship to the underlying construct (Messick, 1989).

Many studies (see above) have been conducted to explore the validity of automated essay scoring. This study contributes to that evidentiary base for the use of IntelliMetric™, specifically within the context of WritePlacer, a direct assessment of writing used for college placement.

**Study Design**

Between 2000 and 2003, more than 250,000 students nationwide took one or more of the three “flavors” of ACCUPLACER® along with one or more multiple choice measures in the ACCUPLACER® program. Most took WritePlacer® Plus or WritePlacer® Texas as WritePlacer® ESL was only introduced in 2003.

Cambell and Fiske (1959) identify the need to collect simultaneous evidence about multiple traits (constructs) measured by multiple methods. They refer to this validation approach as a multi-trait multi-method approach to evaluating the validity of tests. An intercorrelation matrix illustrating the various relationships between several constructs measured in several different ways is created as a source of validity evidence. The relative values provide a trail of evidence; “direct convergent evidence is afforded by the coefficients in the… validity diagonals which should be statistically significant and sufficiently large to warrant further efforts.” (p. 46) Messick goes on to suggest that the values for the same trait measured by different methods should be higher “than for correlations having neither trait nor method in common.”

In short, this approach compares the relationships among multiple traits measured using multiple methods. As captured by Messick (1989), “To demonstrate that a construct is not uniquely tied to any particular measurement method requires the convergence of two or more methods of measuring the construct. To demonstrate that a construct is not redundant vis a vis other constructs requires discrimination among measures of two or more constructs.” (p.46)

One would expect measures of the same trait to be more closely related than measures of different traits. Moreover, one would expect that a measure of a given trait would be more closely related to another measure of that trait than to a measure of the same trait assessed using a different method. While a complete trait by method matrix of scores was unavailable for this study, constructed response and multiple choice measures for writing were available as were multiple choice measures for reading, writing, arithmetic, and college level math.

**Measures and Method**

**Measures.** The following measures were included in the study.
• **WritePlacer® Plus**- A direct assessment of academic writing skills scored on a 6 point rubric
• **WritePlacer® TX**- a version of WritePlacer Plus scored on a 4 point rubric
• **WritePlacer® ESL**- A direct assessment of narrative writing designed for ESL students to assess their level of English skills. WritePlacer ESL is scored on a 6 point rubric
• **ACCUPLACER® Sentence Skills**- A multiple choice measure of grammar, usage and mechanics
• **ACCUPLACER® Reading Comprehension**- A multiple choice measure of reading comprehension.
• **LOEP® Language Use**- A multiple choice measure of grammar and usage designed for ESL students
• **LOEP® Sentence Meaning**- A multiple choice measure of language understanding designed for ESL students
• **LOEP® Reading Skills**– A multiple choice measure of reading skills designed for ESL students
• **ACCUPLACER® Arithmetic**- a multiple choice measure of arithmetic skills including numerical and word problems.
• **ACCUPLACER® College Level Math**- A multiple choice measure of pre-calculus, calculus, and related college level math skills

**Research Questions.** One would expect a valid measure to correlate highly with other measures of the construct and correlate less well with measures of other constructs. Therefore, this study explored the following questions:

**WritePlacer® Plus:**

1. Do WritePlacer® Plus scores generated using automated essay scoring technology correlate highly with other measures of writing and the language arts construct?
2. Do WritePlacer® Plus scores generated using automated essay scoring technology show lower correlations with measures of mathematics than they do with other measures of the writing construct and reading?

**WritePlacer® Texas:**

3. Do WritePlacer® Texas scores generated using automated essay scoring technology correlate highly with other measures of writing and reading construct?
4. Do WritePlacer® Texas Plus scores generated using automated essay scoring technology show lower correlations with measures of mathematics than they do with other measures of the writing construct and reading?

**WritePlacer® ESL:**

5. Do WritePlacer® ESL scores generated using automated essay scoring technology correlate highly with other measures of ESL writing and ESL Reading construct?
6. Do WritePlacer® ESL scores generated using automated essay scoring technology show lower correlations with measures of mathematics than they do with other measures of the writing construct and reading?
7. Do WritePlacer ESL generated using automated essay scoring technology show lower correlations with non-ESL measures of writing and reading than with measures of ESL writing and the ESL Reading construct?
Method. Students took both a WritePlacer® direct writing assessment as well as one or more multiple choice measures as part of ACCUPLACER® Online.

The WritePlacer® scores were correlated with the scores for the four multiple-choice measures. Separate analyses were conducted for the WritePlacer® Plus, WritePlacer® Texas and WritePlacer® ESL programs.

Results

Automated scores for WritePlacer® correlated highly with measures of the same or related constructs (writing and reading) and significantly lower with measures of the math construct (arithmetic and algebra). The pattern of correlations showed the relationships that would be expected if the construct were being appropriately measured. The results were as follows:

WritePlacer Plus:

- **WritePlacer® Plus- Multiple Choice Writing Test (Sentence Skills).** WritePlacer® showed a strong relationship with the multiple choice measure of writing skills (Sentence Skills; Pearson r = .41 p<.001).
- **WritePlacer® Plus- Multiple Choice Reading Test (Reading Comprehension).** WritePlacer® showed a strong relationship with the measure of reading skills (Reading Comprehension; Pearson r = .38 p<.001)
- **WritePlacer® Plus- Multiple Choice Arithmetic Test.** WritePlacer® showed a weaker relationship with the measure of arithmetic (Pearson r = .29 p<.001)
- **WritePlacer® Plus- Multiple Choice College Level Math Test.** WritePlacer® showed a weak relationship with the measure of college level mathematics (Pearson r = .13 p<.001)

WritePlacer Texas:

- **WritePlacer® Texas- Multiple Choice Writing Test (Sentence Skills).** WritePlacer® showed a strong relationship with the multiple choice measure of writing skills (Sentence Skills; Pearson r = .40 p<.001).
- **WritePlacer® Texas- Multiple Choice Reading Test (Reading Comprehension).** WritePlacer® showed a strong relationship with the measure of reading skills (Reading Comprehension; Pearson r = .36 p<.001)
- **WritePlacer® Texas- Multiple Choice Arithmetic Test.** WritePlacer® showed a weaker relationship with the measure of arithmetic (Pearson r = .23 p<.001).
- **WritePlacer® Texas- Multiple Choice College Level Math Test.** WritePlacer® showed a weaker relationship with the measure of college level mathematics (Pearson r = .04 p>.001)

WritePlacer ESL:

- **WritePlacer® ESL- Multiple Choice Writing Test (Language Use).** WritePlacer® showed a strong relationship with the measure of writing skills (Language Use; Pearson r = .41 p<.001)
- **WritePlacer® ESL- Multiple Choice Reading Test (Sentence Meaning).** WritePlacer® showed a strong relationship with the multiple choice measure of sentence comprehension (Sentence Skills; Pearson r = .37 p<.001).
- **WritePlacer® ESL- Multiple Choice Reading Test (Reading Skills).** WritePlacer® showed a strong relationship with the measure of reading skills (Reading Skills; Pearson r = .29 p<.001)
- **WritePlacer® ESL- Multiple Choice Arithmetic Test.** WritePlacer® showed a weaker relationship with the measure of arithmetic (Pearson r = .17 p<.001)
- **WritePlacer® ESL- Multiple Choice College Level Math Test.** WritePlacer® showed a weaker relationship with the measure of college level mathematics (Pearson r = .07 p>.001)

**Discussion**

The results are consistent with what would be expected of an instrument that is effectively measuring the construct it purports to measure. As expected, WritePlacer® showed stronger relationships with multiple choice measures of writing skill and with the reading measure than with the arithmetic or algebra tests. In short, WritePlacer® was more strongly related to other measures of language arts skills than to measures of mathematics skills.

The correlations between the WritePlacer® measure and the multiple choice measure of writing skill was somewhat lower than that which might be expected under ideal circumstances. Unfortunately, the strength of the correlation observed was mitigated by the reduced variance associated with a single measure of writing on a 4 or 6 point scale, with limited variance in the population. Moreover, the reliability of the single item writing measure no doubt attenuates the true relationship.

Establishing the validity of an instrument is an ongoing process; while this study supports the validity of the WritePlacer® assessment, more evidence needs to be collected in support of this claim.
References


### Table 1
**ACCUPLACER/WritePlacer® Correlation Matrix (All measures)**

<table>
<thead>
<tr>
<th></th>
<th>WritePlacer Plus (N)</th>
<th>WritePlacer Texas (N)</th>
<th>WritePlacer ESL (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentence Skills</strong></td>
<td>.40 (57,982)*</td>
<td>.40 (138,058)*</td>
<td>.28 (627)*</td>
</tr>
<tr>
<td><strong>Reading Comprehension</strong></td>
<td>.38 (114,960)*</td>
<td>.36 (145,521)*</td>
<td>.19 (474)*</td>
</tr>
<tr>
<td><strong>LOEP Language Usage</strong></td>
<td>.13 (919)*</td>
<td>.22 (686)*</td>
<td>.41 (1,347)*</td>
</tr>
<tr>
<td><strong>LOEP Sentence Meaning</strong></td>
<td>.11 (650)*</td>
<td>.20 (224)*</td>
<td>.37 (1,354)*</td>
</tr>
<tr>
<td><strong>LOEP Reading Skills</strong></td>
<td>.08 (936)*</td>
<td>.17 (251)*</td>
<td>.29 (2,061)*</td>
</tr>
<tr>
<td><strong>College Level Math</strong></td>
<td>.13 (11,656)*</td>
<td>.04 (20,438)</td>
<td>.07 (385)</td>
</tr>
<tr>
<td><strong>Arithmetic</strong></td>
<td>.29 (81,039)*</td>
<td>.23 (78,823)*</td>
<td>.17 (670)*</td>
</tr>
</tbody>
</table>

* p < .001

### Table 1a
**ACCUPLACER/WritePlacer® Plus and Texas Correlation Matrix**

<table>
<thead>
<tr>
<th></th>
<th>WritePlacer Plus (N)</th>
<th>WritePlacer Texas (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentence Skills</strong></td>
<td>.40 (57,982)*</td>
<td>.40 (138,058)*</td>
</tr>
<tr>
<td><strong>Reading Comprehension</strong></td>
<td>.38 (114,960)*</td>
<td>.36 (145,521)*</td>
</tr>
<tr>
<td><strong>College Level Math</strong></td>
<td>.13 (11,656)*</td>
<td>.04 (20,438)</td>
</tr>
<tr>
<td><strong>Arithmetic</strong></td>
<td>.29 (81,039)*</td>
<td>.23 (78,823)*</td>
</tr>
</tbody>
</table>

* p < .001

### Table 1b
**ACCUPLACER/WritePlacer® ESL Correlation Matrix**

<table>
<thead>
<tr>
<th></th>
<th>WritePlacer ESL (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language Use</strong></td>
<td>.41 (1,347)*</td>
</tr>
<tr>
<td><strong>Sentence Meaning</strong></td>
<td>.37 (1,354)*</td>
</tr>
<tr>
<td><strong>Reading Skills</strong></td>
<td>.29 (2,061)*</td>
</tr>
<tr>
<td><strong>College Level Math</strong></td>
<td>.07 (385)</td>
</tr>
<tr>
<td><strong>Arithmetic</strong></td>
<td>.17 (670)*</td>
</tr>
</tbody>
</table>

* p < .001
Appendix A

IntelliMetric Technology

“...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 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. 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 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. 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.

**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.
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 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”, 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 to 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.

**Non linear.** 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 students’ 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

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.** 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: Pre processing.** After the information has been received in electronic form, IntelliMetric™ prepares the information for further analysis. This preprocessing stage makes sure that 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 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 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:

- **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.
References


