5

TAKING A SPIN ON THE INTELLIGENT ESSAY ASSESSOR

Tim McGee

The following narrative recounts an experiment I performed upon a particular essay-scoring machine, the Intelligent Essay Assessor (IEA), that was first brought to my attention by Anne Herrington and Charles Moran's 2001 *College English* essay "What Happens When Machines Read Our Students' Writing?" As a writing program administrator, I was experienced in the progressive waves of writing assessment historically performed by human readers and well versed in many aspects of computer-assisted instruction. At the time of my experiment, however, I was still a neophyte in the area of automated essay scoring. Nevertheless, despite serious misgivings about my qualifications in the arcane realm of artificial intelligence, I took to heart Herrington and Moran's call to learn about these programs "so that we can participate in their evaluation and can help frame the debate about the wisdom of their use in our own institutions" (484–85).

This account of my experiment with IEA and how it helped me frame the debate about machine scoring, first to local colleagues and later at the NCTE conference in Baltimore, represents both a report of my research and a story about how English teachers (and other mere humanists) might respond to corporate vendors of increasingly sophisticated programs in the technically bewildering arena of automated essay scoring. Consequently, in addition to recounting my method, results, and conclusions, I have included certain historical and biographical material to help the reader understand how my practice was informed by theory (of both textual analysis and writing assessment) and motivated by site-specific relationships of knowledge and power.

NOT YOUR FATHER'S SCORING MACHINE

Unlike the scoring machines that are aimed specifically at the needs of large-scale placement assessment, IEA is pitched as a "new learning

tool, useful in almost any subject" that promises to ease the burden of assigning writing across the curriculum and purports to measure the "factual knowledge" displayed in student essays. IEA is just one of several products marketed by Knowledge Analysis Technologies, whose Web site included the tagline "Putting Knowledge to the Test" and promised "[m]achine learning technology that understands the meaning of text."[1] Compared to the average academic sites, many of which had yet to exploit the graphical possibilities of the Web, Knowledge Analysis Technologies' home page was as visually enticing as the most polished commercial and entertainment Web sites. At the same time, the easily navigated site provided the scholarly apparatus one normally associates with academic research, including impressive lists of publications by the company's principals and full-text access to several white papers. (The home page has since been toned down visually, the heavily Photoshopped montage of learners—including smiling children, a female soldier, and a U.S. flag—now replaced by a file-folder navigation bar and a color scheme to match that of the new corporate parent, Pearson Education.) The rhetorical sophistication of the original site was equally impressive, as the content and tone of the promotional copy aimed at military clients contrasted noticeably with those portions of the Web site where the implied audience was college professors. However, most stunning of all were Knowledge Analysis Technologies' invitations to "[i]magine intelligent Internet technology . . . that understands the meaning of written essays, evaluates them and provides feedback as accurately as a professional educator or trainer." Particularly amazing were the following claims:

> IEA is the only essay evaluation system in which meaning is dominant. It measures factual knowledge based on semantic content, not on superficial factors such as word counts, punctuation, grammar or keywords. IEA also provides tutorial commentary, plagiarism detection, and extensive validity self-checks. And it does it right now—not in days or weeks. (Knowledge Analysis Technologies 2001)

What made these claims stand out were the bold assertions about understanding meaning, both in light of conventional wisdom among compositionists and in comparison to the far more cautious claims of competing vendors. The conventional wisdom had been succinctly stated ten years earlier by Fred Kemp when he wrote that "computers can process text in only the most superficial of senses; computers cannot grasp the meaning in the text" (1992, 14). While it was possible that great

leaps in natural language processing had been made in the intervening decade, Knowledge Analysis Technologies' claims still sounded outlandish when compared to ETS pronouncements about e-rater, the scoring engine behind Criterion and other products, which scrupulously avoided even claiming that e-rater "read" essays, much less "understood" them. As a potential adopter of products that seemed likely to help teachers and students with the important work of assessing writing, I was interested in interrogating Knowledge Analysis Technologies' claims.

INSTITUTIONAL SETTING

I was interested not just in the sense of piqued curiosity but, as an administrator cum teacher/scholar, I was an interested party in the shifting interrelationships among teaching, research, and commerce now found in several areas of computer-mediated communication. That interest was complicated by an accident of geography that put my institution in a particularly cozy relationship with ETS; while not in the vanguard of embracing instructional technology, the college was poised to adopt, or at least try, machine scoring of some student essays.

Located fifteen minutes from the national headquarters of ETS, the college had numerous faculty members who regularly worked as evaluators and consultants in several content areas, while the school generated considerable income every year by renting out blocks of classrooms for mass scoring of various tests. The impacts were not strictly financial, as the employment opportunities also yielded a familiarity with holistic essay scoring that extended well beyond the disciplinary borders of English and composition. For example, when the School of Business requested a workshop to help its faculty integrate writing into their curricula, I learned that some business professors were already using holistic scoring guides lifted from a Graduate Management Admissions Test essay-scoring session by one of their colleagues.

The college's interest in investigating machine scoring had already found expression from various corners, including the dean of Academic Support, the director of the Economic Opportunity Fund Program, and the Writing Assessment Committee of the School of Business. It was the last group (formed as part of the pursuit of Association to Advance Collegiate Schools of Business accreditation) that actively pursued machine scoring, as the business faculty sought a mechanism for implementing a value-added assessment of their students' writing skills that would meet their own quantitative notions of reliability and validity while also appearing objective to outside evaluators.[2] This led to the

purchase of enough access to Criterion for a pilot study in which a portion of incoming business majors wrote to a GMAT-style "issue" prompt.

As a result, I had some firsthand experience with a scoring engine that relied, as Knowledge Analysis Technologies dismissively put it, upon "superficial factors such as word counts, punctuation, grammar or keywords," and I held the quaint notion that an evaluation system that claimed to get at "meaning" might not want to abandon punctuation, much less grammar. In other words, I was willing to grant limited possibilities for machines that performed automated essay scoring the old-fashioned way: counting, measuring, and using keywords and parsers for recursive syntactical analysis to compare a new sample essay to a large batch of essays previously scored holistically by trained human readers.

Admittedly, that willingness bespeaks a certain unreconstructed New Critical approach to textual analysis that many (myself included) now find theoretically problematic when talking about serious analyses of texts, including student texts. However, the assembly-line analyses of timed impromptu essays written to a "general knowledge" prompt by students under duress, with no recourse to the usual parts of their composing process (much less such aids as dictionaries or peer critics) is already such a constrained response to a rather inauthentic text-production event that the analytical approaches designed to remedy the severe limitations of New Criticism need not be invoked. In other words, given the severe limitations of what a short impromptu essay test allows students to display,[3] an analytical approach that assumes the meaning and value of a piece of discourse is discernible by an examination of the text itself is not theoretically inappropriate. While fully agreeing with Herrington and Moran's (2001) conclusions about other harms that machine scoring does to the entire project of rhetorical education, I believed that the latest generation of scoring engines could, in fact, replicate the scores given by humans in the relatively restricted domain of large-scale placement assessments.[4]

But that is a far cry from accepting the claim that a scoring engine "understands the meaning of written essays." And given that humans would be hard pressed to understand an essay without relying, to some degree, on "punctuation, grammar or keywords," I was thoroughly mystified about how IEA could possibly do so. Neither the promotional copy of the Knowledge Analysis Technologies site nor the teacher-friendly account provided by Herrington and Moran did much to demystify how the artificial intelligence behind IEA actually worked. Here is what

the promotional portions of Knowledge Analysis Technologies' site said about IEA's approach:

> The Intelligent Essay Assessor uses Latent Semantic Analysis, a machine-learning algorithm that accurately mimics human understanding of language. This patented and proprietary technology is based on over 10 years of corporate and university research and development. IEA analyzes the body of text from which people learn to derive an understanding of essays on that topic. The algorithm is highly computer intensive, requiring over a gigabyte of RAM, which is why IEA is offered as a web-based service. (2001)

Herrington and Moran call IEA "quite an interesting product" and provide two pages of eminently readable explanation that begins as follows:

> IEA derives from Thomas Landauer's work on what he termed "latent semantic analysis." Latent semantic analysis (LSA) is, briefly and for our purposes, based on the assumption that there is a close relationship between the meaning of a text and the words in that text: both what these words are and how these words are related to one another in the space of the text. Landauer and his group are not talking here about lexical or grammatical relationships but about spatial relationships: what words the text includes and in what spatial relationship to one another. For their purposes, lexical and grammatical relationships are irrelevant. (2001, 491)

They go on to explain that Landauer and company posited the existence of "vast numbers of weak interrelations" in some domains of knowledge and the ability to describe them mathematically. This, in turn, allows a machine "to measure whether someone has learned something or not by looking at the text that person produces and seeing whether this text contains some of the 'vast number of interrelations' that are characteristic of the material that was to have been learned" (491). The focus on learning content is peculiar to IEA because, unlike the machines marketed primarily as aids to placement assessment, IEA promises to help teachers and learners by evaluating essays based on what their authors appear to know about a topic.

METHOD

Intrigued by the prospects of this seemingly revolutionary approach to machine scoring, I wanted to design an experiment that would give me a better sense of how IEA actually worked. I modeled my method upon that of Herrington and Moran, who had submitted multiple drafts to the machines, watched the scores change, and then asked what the ratings

seemed to indicate about how the machines "read" the essays and what criteria were operating (2001, 490). In the end, I wanted to compare IEA's notion of "meaning" with my own. That led me to consider what I meant by "meaning," not in a broad philosophical sense, but in the relatively restricted domain of student essays written to specific prompts.

Rarely, when reading impromptu essays, had I felt compelled to decode ironies, ponder obscure cultural references, or interpret subtle uses of symbolism. The textual ambiguities I regularly encountered in impromptu essays rarely seemed intentional, productive, or fodder for deconstructive performance; rather, they usually appeared to be the results of imprecise word choice and careless syntax. I concluded that making face-value meaning out of impromptu essays is an interpretive process that relies primarily upon lexicon, syntax, propositional content, and the arrangement of ideas. In effect, I had no reservations about granting the machines ample ground on which to perform admirably in the analysis of multiple textual features that, in my estimation, contribute heavily to face-value meaning. Furthermore, having taught composition mostly to well-prepared first-year college students, I held the view that most were able to compose legal sentences in decent paragraphs but not yet skilled in global text arrangement, especially when writing arguments (as opposed to narratives, reports, or expositions of processes). As a result, I was of the opinion that arrangement exerts considerable influence as a higher-order source of meaning, especially in student essays.

Consequently, when I began my experiment, I had some positive expectations about the potential for an analytical approach that depended in part upon how "words are related to one another in the space of the text." My intention was not to trick the machine into awarding high scores to meaningless gibberish, but rather to make some calculated revisions to texts that the machine purportedly scored well so I might consider what the changed scores told me about how IEA was "reading" these essays and what criteria were operating for determining meaning.

While Knowledge Analysis Technologies' claims about IEA understanding meaning might seem to have invited something like a Turing test,[5] my aims were considerably more modest. Holding no illusions about IEA deserving to be deemed intelligent based upon any dialogue with a user, I was simply attempting to get a fix on Knowledge Analysis Technologies' definition of and criteria for meaning. I proposed to accomplish this by analyzing what features of a text appeared to affect the evaluations produced by a "system in which meaning is dominant." I

also had a desire to operate upon something like the principle of charity, submitting only essays that might meet the criteria for what the some of the test vendors call a "good faith effort."

In a panel session at CCCC 2001 titled "Challenging 'E-rater': Efforts to Refine Computerized Essay Scoring," Mary Fowles of ETS recounted how their researchers used various tactics to trick their scoring machine and then used those results to refine the program. While she and other representatives of ETS have admitted that it is possible to design essays specifically aimed at tricking the machines into awarding top scores to texts that no human would rate highly, they contend that a fair assessment of the machine's reliability and validity depends upon the submission of essays that are like ones that real students would actually submit, what they refer to as "good faith efforts."[6] However, any requirement to limit the revisions of their sample texts to ones that possessed some degree of verisimilitude to actual student texts would represent a substantial restriction upon my efforts to quickly ascertain what features contributed to meaning. Furthermore, such a restriction would seem to turn my experiment back in the direction of a sort of reverse Turing test, as if I were attempting to ascertain when IEA knew that the submission was not from a real student. So, I opted to look at the sample essays IEA offered and try to determine what specific characteristics of each essay seemed most integral to its meaning.

RESULTS OF THREE SPINS ON THE MACHINE

The Knowledge Analysis Technologies Web site provided unfettered access to the "Intelligent Essay Assessor™ Demonstration Page," which included five different "content-based essays" that visitors could experiment with. Each of the five was identified by subject, topic, and grade level. These were the choices:

- Biology: Function of Heart and Circulatory System (College Freshman)
- Psychology 1: Attachment in Children (College Freshman)
- Psychology 2: Types of Aphasia (College Freshman)
- Psychology 3: Operant Conditioning (College Freshman)
- History: The Great Depression (11th Grade High School) (KAT 2004a)

The instructions give the visitor the choice to "compose your own essay or use one of the sample essays provided" and include, for each

essay, a prompt that requests (with varying degrees of specificity) a certain response from the writer. The prompt for the "Function of the Heart" essay makes the following request:

> Please write down what you know about the human heart and circulatory system. Your essay should be approximately 250 words. We would like for you to be as specific as possible in discussing the anatomy, function, and purpose of the heart and circulatory system.

The IEA demonstration page included three different essays in response to the "Function of the Heart" topic, each one scored on a 1 to 5 scale in four categories: overall, content, style, and mechanics. The best of their three sample essays (biology sample 1) scored 4 overall, receiving 4 for content and 3 each for both style and mechanics. This sample starts with a functional definition ("The heart is the main pump in the body that supplies the rest of the body with oxygenated blood by way of the arteries") and then proceeds through an orderly exposition that includes the replacement of oxygen in the blood by CO_2, traces the path of the blood through the veins to the heart and lungs, and concludes "now the blood will be pumped to the rest of the body and the cycle begins again." As I attempted to determine how I made meaning of this particular essay that explained a biological process, I decided that I was relying heavily upon a combination of lexicon, syntax, and sequence, especially in terms of the various techniques used to foster cohesion from one sentence to the next.[7]

Experiment 1

I was struck by the highly sequential nature of the exposition and imagined that the aptness of the particular sequence the author had chosen had considerable bearing on both the correctness of the content and the global coherence of the essay. I wondered what effect changing the sequence of the sentences might have on the essay's score. I assumed that such a change would have no effect on the essay's mechanics score, but should have some effect on its style score, and wondered just how large an effect changing the sequence of the sentences might have on the essay's content score. So, I took biology sample 1 and, leaving each individual sentence unchanged, reversed the order of its thirteen constituent sentences, so the first sentence becomes the last and vice versa. The result is a rather peculiar text that doesn't actually describe the heart and lungs working opposite of the way they really do. Rather, the effect is more like that of the movie *Memento*, in which each individual section of narrative

runs chronologically but the narrative as whole runs backward.

Hence, the revised essay begins as follows:

> The left ventricle is the most muscular of the heart because now the blood will be pumped to the rest of the body and the cycle begins again. The blood is then pumped into the left ventricle through the left atrioventricular valve. The blood, now oxygenated, goes back to the heart by way of the pulmonary vein and then into the left atrium.

The meaning of the individual sentences is unchanged, but the assembled whole has suffered a substantial reduction in both cohesion and coherence, not to mention factual accuracy. I was surprised (and disappointed) when IEA awarded the exact same score to the fully reversed essay as it had awarded sample 1. Someone with a better understanding of latent semantic analysis might have guessed that the reversed sample 1 would receive a score identical to the original sample 1 because, to the Intelligent Essay Assessor, the two are the same essay. However, to mere mortals who rely upon cohesion, coherence, sequence, and arrangement as ways to make meaning of written discourse, sample 1 and reversed sample 1 are radically different essays, with reversed sample 1 providing a substantially less meaningful exposition of the function of the heart and circulatory system. I therefore concluded that global arrangement is not part of IEA's notion of "meaning." Even more disconcerting was the realization that neither cohesion nor coherence (in the senses used by Joseph Williams in *Style* [2000]) had any impact on "meaning" as that term is used by the producers of IEA.

Experiment 2

Based upon Knowledge Analysis Technologies' claim that IEA "measures factual content," I attempted to see how a change to the factual content of an essay might alter its score. Of the various samples available on the IEA demonstration page (2004a) the history essays on the Great Depression seemed the ones most chock-full of factual content. The prompt for the history essay is not nearly as specific as the one for biology, asking simply, "Please write a structured essay on the 'Great Depression' and the 'New Deal.'"

Again, history sample 1 received the highest score, getting a 5 overall, with 5 for content and 4s for both style and mechanics. At 564 words, it was the longest essay in the IEA demonstration page and seemed the best candidate for attempting to determine how IEA's measurement of factual content affected its analysis of meaning. The most

straightforward way I could imagine altering the factual content of the essay was to simply reverse the truth value of many of its propositions. Where the original essay wrote "was," I substituted "was not." "Biggest" became "smallest," "before" became "after," and "start" became "end." For example, here is the beginning of original history sample 1:

> There were many problems facing the nation in 1938, following the stock market crash in 1929 and in the midst of Franklin D. Roosevelt's New Deal. Roosevelt, a moderate, attempted to combat the system of rising tariffs, expand opportunity in business for the independent man, reestablish foreign markets for America's surplus production, meet the problem of under consumption, distribute the nation's wealth and instigate a level playing field in America.

The revised history sample 1 begins as follows:

> There were few problems facing the nation in 1929, following the stock market crash in 1938 and at the end of Franklin D. Roosevelt's New Deal. Roosevelt, a radical, attempted to promote the system of rising tariffs, diminish opportunity in business for the independent man, end foreign markets for America's surplus production, meet the problem of over consumption, centralize the nation's wealth and instigate a tilted playing field in America.

This process was continued throughout all twenty-four sentences of the original sample. Clearly, the factual content in the revised sample 1 is markedly different from that in the original sample 1. A machine that somehow "measures factual content" ought, it would seem, to come up with a different measurement, unless, of course, it measured only the amount of factual content, in which case, the revised sample 1 might measure up to the original. But surely, the meaning has changed substantially. To "diminish opportunity in business for the independent man" means the opposite of to "expand opportunity in business for the independent man."

Frighteningly, IEA awarded the same high score of 5 (with all the same subscores) to the revised sample 1, despite the fact that it is as factually inaccurate as could be while still being an essay on the topic of the Great Depression and the New Deal.[8] So, unlike this humanist's definition of meaning, on which such pedestrian notions as the logical denotation of a phrase have considerable bearing, IEA's notion of meaning appears to exist quite independent of any relationship to factual accuracy. And yet, amazingly, Knowledge Analysis Technologies claims that IEA "measures factual content."

Experiment 3

Having acted in what I considered to be good faith on my first two efforts and beginning to feel like a bit of a chump for having believed that this machine could, in fact, isolate something in a text that bore some relationship to what normal people consider to be the text's meaning, I was now ready to submit something outlandish, but perhaps not quite as outlandish as Knowledge Analysis Technologies' claims about IEA understanding meaning and measuring factual content.

I chose "Psychology 2:Types of Aphasia" and endeavored to find out if there was, in fact, anything that IEA did not find meaningful. I was looking to see if there was a bottom discourse that IEA would not accept as a meaningful text. The actual prompt for "Types of Aphasia" I considered to be one of the best and most interesting on the IEA demonstration" page, because of its high specificity and potential for eliciting somewhat authentic displays of knowledge and understanding. Here is the prompt:

> After a mild stroke, Mr. McGeorge showed some signs of aphasia. What pattern of symptoms would lead you to believe he had suffered damage primarily in: (a) Broca's area, (b) Wernicke's area, (c) the angular gyrus? (2004a)

This collection of samples was scored on a 10-point scale, with sample 1 receiving a 7 overall, with 7s for content and style and a 6 for mechanics. On this sample revision, I preserved much of the original vocabulary and maintained most of the sequence, while turning the diagnosis itself into nonsense, complete with multiple cases of mangled syntax. Here is the entire original sample 1:

> To detect the effects that Mr. McGeorge's stroke had I would conduct several experiments testing his ability to communicate. If he had trouble verbalizing words I would be alerted that his Broca's area of the left frontal lobe was damaged. However, if he could not even comprehend the meaning of a word that would indicate damage to his Wernicke's area of the left temporal lobe. Finally, if Mr. McGeorge could not even "see" the words in his head, or understand writing, I would conclude he had damaged his angular gyrus located in the occipital lobe. (It is assumed that Mr. McGeorge is right-handed with his speech center being the left hemisphere).

And here is the revised sample 1 that I submitted to IEA:

> To effect the detects that Mr.stroke McGeorge had I would several conduct experiments testing ability his communicate to. If he had trouble verbalizing the left frontal lobe I would alert Tom Broca that his communicate was damaged. However, if he could not even meaning the comprehend of a word that would indicate damage his to area Wernicke's the of left lobe temporarily. Finally, if Mr. McGeorge could not even "pronounce" the words in his mouth, or understand the meaning of finally, I would fasten his angular gyrus to his occipital lobe. (It is assumed that Mr. McGeorge is even-handed with his speech center being on the far left wing).

I was pleased to discover that IEA did not award this gibberish the same score as the original, and did, in fact, reduce its content score. That indicated to me that the machine really did process submitted texts in some fashion—apparently the lights were on and somebody was home. Unfortunately (or fortunately for my purposes, which had undergone some revision in the course of the experiment), IEA awarded the revised sample 1 the same overall score, because the one content point it lost (slipping from a 7 down to a 6) was balanced by the one point it gained in the area of mechanics (rising from a 6 to a 7). This caused me to wonder what the makers of IEA could possibly mean by "mechanics" if the revised sample 1 was mechanically superior to the original.

However, by that time I had seen enough to draw two conclusions: the meaning of "meaning" that Knowledge Analysis Technologies was using in its claims about IEA was nothing like the conventional meaning of that word as used by laypeople, humanists, compositionists, or even such esoteric groups as philosophers of language. The meaning of a text that latent semantic analysis actually gets at, if in fact it gets at any, is so far removed from any notion of meaning that anyone assigning writing to students would be employing that it appears to render the claims that Knowledge Analysis Technologies was making about IEA's analytical abilities patently false. Latent semantic analysis does appear to do something, but whatever it does appears to be wildly unsuited to the scoring of student essays. Whatever subtle information latent semantic analysis may yield, the Intelligent Essay Assessor's performance on the three sample essays was seriously at odds with and far inferior to the results of blatant semantic analysis, or the meaning that a mere mortal might make from those sample texts.

PRESENTATION OF FINDINGS

Shocked as I was by the inadequacies of IEA for evaluating student essays, and appalled as I was at the thought that this product was being marketed to high school and college faculty as an appropriate tool to aid in the integration of writing "in almost any subject," I felt compelled to share my views with my colleagues, first at a collegewide Teaching and Learning with Technology workshop and later at the 2001 NCTE conference. At both venues, the reenactment of my experiment was met by a mixture of dropped jaws and howling laughter. So, for two audiences, with a total number of perhaps fifty souls, I was able to demonstrate that one particular approach to automated essay scoring was unlikely to be as useful as the vendor's promotional copy would lead potential adopters to believe. Meanwhile, stories about IEA kept appearing in the mainstream media, telling millions of people what Knowledge Analysis Technologies said their product promised to do.

At my own institution, there was never any likelihood that IEA was going to be adopted widely, and even the pilot use of Criterion turned out unsuccessfully, as too many of the first-year students at that selective college scored near the top of the 6-point scale for ETS's machine to serve the value-added purposes that the School of Business had hoped to use it for. But I shudder to think how many high school and college students have already had their rhetorical education impacted by the introduction of Knowledge Analysis Technologies' IEA into the curriculum.

As I mentioned, since being purchased by Pearson Education, Knowledge Analysis Technologies' Web site has been toned down considerably, both visually and in terms of its claims about understanding meaning. However, four years after my original experiment, IEA still works as abysmally as it did in 2001; the scores it awards to the three revised samples are unchanged from those it coughed up four years ago. But now it has the corporate backing of Pearson Education, a company that many educators associate with an outstanding collection of composition and rhetoric titles from what used to be the publishers Addison-Wesley, Longman, and Allyn & Bacon, but are now Pearson "brands." The combination of deep corporate pockets, the credibility that attaches to Pearson's stable of authors, and the marketing ploy of bundling ancillary Web resources with textbook adoptions seems likely to spell huge increases in the deployment of IEA upon unsuspecting students hoping

for meaningful responses to drafts and polished essays, a prospect I find both frightening and depressing.

MORAL OF THE STORY

My experience with IEA began with Herrington and Moran's essay in *College English* (2001) and their call to learn about the scoring machines and to participate in the debate surrounding them. I quickly learned that despite my inability to engage in an informed debate about the merits of the artificial intelligence behind IEA, it was really quite easy to demonstrate that latent semantic analysis, at least as it is embodied in IEA, cannot be trusted to score student essays well. In effect, it may have been my innocence in the realm of artificial intelligence that led me to this emperor's new clothes sort of revelation. I concluded that IEA represents a form of automated essay scoring that no conscientious educator would unleash upon students wanting meaningful evaluation of their writing. In the end, my experience did help me frame the debate in my own institution, and I hope that the presentations of my findings can help others to do the same.