6

ACCUPLACER'S ESSAY-SCORING TECHNOLOGY

When Reliability Does Not Equal Validity

Edmund Jones

Placement of students in first-year writing courses is generally seen as a time-consuming but necessary exercise at most colleges and universities in the United States. Administrators have been concerned about both the expense and inconvenience of testing, about the validity of the tests, and about the reliability of the scorers. Over the past decade, computer technology has developed to the point that a company like ACCUPLACER, under the auspices of the College Board, can plausibly offer computer programs that score student essays with the same reliability as expert scorers (Vantage Learning 2000). Under this system, schools need hire no faculty members to score essays, and students can arrange to be proctored off-site; thus placement testing becomes far more convenient without increasing costs. In fact, for these very reasons, Seton Hall University currently uses ACCUPLACER to aid in placing students in College English I and basic skills courses. On the middle school and high school level as well, classroom teachers appear willing to use computer ratings to help rate students or to supplement their feedback on students' writing (Jones 1999).

However, for reasons both theoretical and pedagogical, some in the discipline of composition have questioned the appropriateness of using computers to score writing. In a critical discussion of machine scoring in *College English* in 2001, Herrington and Moran raise several concerns. They wonder about the effect that writing for a computer instead of a human being will have on composing an essay. And they believe that "an institution that adopts the machine-reading of student writing sends its students two messages: human readers are unreliable, quirky, expensive, and finally irrelevant; and students' writing matters only in a very narrow range: its length, its vocabulary, its correctness," or its ability to conform to what a computer can measure (497).

Perhaps living with less theoretical concerns, professional testing administrators at New Jersey colleges have generally embraced

ACCUPLACER (Kozinski 2003). And they are happy about WritePlacer *Plus,* the direct writing-assessment component of ACCUPLACER, precisely because it *is* reliable. Reliability refers to the ability of a scorer, whether human or machine, to give the same score consistently to essays of similar quality and for one scorer to give the same scores as another scorer. When two humans score an essay, there is always the possibility that they will have somewhat different criteria and, consequently, score the essay differently—unless, of course, they are carefully trained under controlled circumstances. This training takes time and money. If a computer can be trained to score essays, on the other hand, reliability problems should disappear. IntelliMetric, the proprietary electronic essay-scoring technology developed by Vantage Technologies, always scores the same way. As for human-computer interreliability ratings, according to ACCUPLACER its computers agree with human scorers within one point between 97 percent and 99 percent of the time (Vantage Learning 2000).

But the directors in the writing program at Seton Hall wondered if, despite high marks on reliability, IntelliMetric lives up to Vantage Learning's claims for construct validity.[1] That is, we wondered if the computer evaluates what we think it is evaluating. In this concern about validity over reliability we are not alone. Powers et al. (2002) explain that computers will always agree with each other, thus being reliable, but that they may be programmed to focus on a restricted number of criteria for evaluating essays (409). While Herrington and Moran (2001) raise questions about the validity of writing for a nonhuman audience, they also wonder whether the computer can be trusted to evaluate some of the nuances of writing. On the local level, Nancy Enright and I, both directors in the English department at Seton Hall University, had informally come to the conclusion that WritePlacer *Plus* rewards essay length out of proportion to its value. Like Herrington and Moran, however, we couldn't go beyond developing hunches about the validity of the scoring itself because we hadn't systematically analyzed data from student placement essays. We could only suspect that length and mechanical correctness matter to IntelliMetric. One might wonder, though, why not just write to Vantage Learning to ask how the computer goes about scoring the essays? The answer: proprietary information is not divulged by companies that have created software to evaluate writing. As a result, we needed to work with the only evidence we had about how the computers worked: computer-generated essay scores.

Herrington submitted an essay and two revisions—one to improve the original and one to weaken it—to investigate IntelliMetric's powers of discrimination. This approach is logical enough, but Vantage Learning might argue that she wasn't writing the way students actually write essays. After all, IntelliMetric "learns" how to score essays by digesting actual students' essays along with scores given by expert human scorers. In order to address this potential criticism, I thought it important to identify anomalies in actual placement-test essay scores before performing any experiments. Two other instructors at Seton Hall University and I reviewed 149 essays submitted by incoming freshmen Seton Hall students from the summers of 2002 and 2004. We grouped essays by score and then read them, thus deliberately norming ourselves against WritePlacer *Plus*'s holistic scoring system. (Although it would be possible to critique holistic scoring in general, I wanted to examine ACCUPLACER on its own terms—to critique machine scoring in a way that would speak most effectively to the majority of institutions which, for better or worse, accept holistic scoring.) Our method allowed us to readily identify those essays that seemed significantly worse than or better than others in that score group. From this collection of anomalously scored essays, we searched for patterns: which types of essays posed problems for WritePlacer *Plus*? In a preliminary study (Jones 2002), I was able to find a pattern among essays that scored high but seemed weak: they tended to be long. I found a second pattern among essays that were mechanically correct and well developed but seemed weak: they had awkward phrasings that didn't look like English. I expected these patterns to enable me to identify hidden criteria and to identify problems that are invisible to the computer.

Once these hidden criteria were identified, I planned to enter doctored essays, always starting from actual student originals. For example, after Nancy and I suspected that essay length was overvalued, I chose two essays that were each awarded a 6 (out of 12) by WritePlacer *Plus*, appended one to the other, and resubmitted them as one essay. The resultant score? A 9. The computer did not seem to recognize the incoherence that must result from such an operation but apparently rewarded length as a value by itself. Others have used the method of identifying anomalous scorings (Roy 1993), but none, as far as I know, have systematically investigated the ability of the computer to discriminate according to specific criteria by submitting doctored essays.

THE ANOMALOUS ESSAY

WritePlacer *Plus* scores essays based upon five criteria: focus, development, organization, sentence structure, and conventions. I will not challenge the writing construct behind these criteria, though it would be possible to do so. Similar criteria are used in many holistic assessment rubrics. Certainly a somewhat different construct lies behind the WPA Outcomes Statement for First-Year Composition (Council of Writing Program Administrators 2000), which focuses on rhetorical knowledge, critical thinking and reading, writing processes, and conventions. The WritePlacer *Plus* criteria are, perhaps, appropriately narrow because of the decontextualized nature of the writing assignment students face when taking a placement test. How can any reader—human or machine—really consider audience, for example, or consider what students know about multiple drafts? My interest is to take on WritePlacer *Plus on its own terms.* If it doesn't work on its own terms, it cannot meet the minimum standards for validity.

WritePlacer Plus scores range from 2 to 12, though in practice, at least at our institution, there are no 2s or 3s. The great majority of institutions use 8 or 9 as a cutoff point for their college English courses, according to Suzanne Murphy (2004), an associate director at ACCUPLACER. Certainly, an essay that WritePlacer *Plus* scores as a 9 should be an acceptable one. Here is a typical example, in response to a prompt about the advisability of working a second job or overtime:

> I believe working full time or having a second job is too stressful for a person. Many people who are in this position never have time for themselves and their families. Their social life always revolves around their employees and their is no change in their lives. The overwork can also disturb a person academically and physically due to the lack of exercise.
>
> To become better people, many of us need to relax and to take time out for ourselves and with our families. It is very important to spend quality time with the people we are closest to because they are the ones who will help us with all your problems. People with two jobs, and who work overtime have that type of companionship but can not take advantage of it because of their jobs. This can result in many negative effects like depression and not a very good social life.
>
> Having the same, overworked routine everyday is not the ideal life. People who are in this position are very bored of their lives due to the same working habits everyday with the same people. Many overworked people also become cranky and moody because of their jobs and are not

> polite to the customers. This can make matters worse because they do not work with passion, they work because they have to. They also envy the people around them that are enjoying their lives and are partying on friday and saturday nights!
>
> People who work full time, or have two jobs are never in shape because they never receive the excerise they need. They can not concentrate on their health as much due to all the work on their mind. They are also disturbed academically due to the small time frame they put into their education. I have seen many cases where overworked students have gotten failings grades when they were capable of higer scores. Jobs can really divert a persons attention from the important aspects in their lives, which is wrong.
>
> As a student, I believe everyone should get an education and have one job that fullfills their life. This way, there will be time for a good personal and social life. It is important for a person to give time to everything and to live life to the fullest. Everyone is here in the world to enjoy, not just to work. In conclusion, I believe that jobs are important, however, to a certain extent because it is more important to enjoy life. (415 words)

There is a simple but recognizable organization to the essay: introduction, thesis with predictor statements, three body paragraphs, and conclusion. Sentences are generally well constructed and grammar errors don't interfere with understanding. The author implicitly acknowledges that examples can be a useful thing, even if there is no recognition of the value of multiple perspectives or of complexity. But then, given testing conditions and training in writing the five-paragraph theme, it is not surprising that virtually no student exhibits these latter qualities. In any case, I had no problem placing this student in the College English class at our school, where the average verbal SAT score that year was 531.

In 2002 we found an essay whose WritePlacer *Plus* score, also a 9, stunned us—so much so that we asked ACCUPLACER what might account for such an apparent error in scoring—but before going any further, please read the essay for yourself. The writing prompt asks students to judge which form of technology has had the greatest impact on our society.

> Technological advances in the world today or in our daily lives have hit in the business world and the home itself. By bringing technology to high standards has in a way sped up the way of life of how it used to be lived. The most largest affect that anyone one man or woman has used in technology is the computer.

> The computer itself has been the key or the base of this change in life. The usage of the computer has probably excess in millions from 1992 to the modern day we live. Kids these days of me asking how many computers they have in one household is brought to me that there families out there that carry more then one.
>
> However, the computer is just the base like I said, because what the computer carries is more to push us up the technology ladder. Now there are the word processors that have take our typing machines and just toss them in the garbage. Then, the microsoft excel programs which have shaped charts then need to be graphed for business and school. There is also the new programs like adobe pagemaker which is the process of making cards and brochures and using scanned pictures, also, microsoft power point what is basically makign your own power point film, by using slides.
>
> Although, these programs have consisted in the building of a high corporate ladder of technology, nothing evens out with the world of the internet. The internet has shaped everyone in the world into some type of character on the internet. The internet has been the faster things to catch on to by anyone one invention. Even the process has always been going on in the U.S. government but when given to the public it took off like a rocket. It's like faster then anyone other library in the world where tons and tons of information are being transfered back and forth. Even though there is the msn chattings and the aol chattings with just surfing the net there is this great monster coming at the internet at full speed.
>
> That monster is the online gaming that is going on in the world and the leading forfront is the games of Quake and Counter-strike. From the United States to Turkey to England these games are being played. Of course I am one of them and from other experiences and my own it is addictive.
>
> The feelings that I have to the new inventions that have come out in the past quarter of a century have no affect like the computer itself and the power of the net. I really can't think of anything else in the near future that could top this phenomenom that has been growing so fast. But if so by that time we should be living in the planet Mars and my grand kids learn about these computer and internets that we used. (486 words)

Although this student, Carl,[2] shows real enthusiasm for his subject, does generally focus on how computers have sped up our lives, and presents some evidence in support of the thesis that computers are the most important technological advance today, his prose is very tough to plow through. The first paragraph is not more egregious in its language than other paragraphs, but it provides significant insight into the magnitude of the sentence-level problems this student has. I will go into some depth to make the case that the language problems are not insignificant.

> Technological advances in the world today or in our daily lives have hit in the business world and the home itself. By bringing technology to high standards has in a way sped up the way of life of how it used to be lived. The most largest affect that anyone one man or woman has used in technology is the computer.

If you were to read this in Word, you might be surprised to discover that there is only one grammar error noted: the double superlative, "most largest." In fact, this is the least troublesome problem in simply getting through the prose. The word "hit" in the first sentence stopped me briefly. It seems an odd word to use; we might expect "occurred." However, the word "hit" might have been more easily negotiated if there weren't redundant phrases signifying where the technological advances are occurring. The sentence would be clearer if it read as follows: "Technological advances have hit the business world and the home itself." The syntax confusion in the second sentence is profound. It begins with a prepositional phrase used as a subject and ends with the strange interjection of the phrase "of how it" that mars what would have been more comprehensible if left out: "sped up the way life used to be lived." The last sentence has, in addition to the double superlative, the following nonsensical kernel sentence: "Any one man or woman has used the most largest affect in technology." A possible revision of this paragraph, applying principles of syntax, concision, and correctness, would look like this:

> Technological advances affect us daily in the business world and in the home itself. Bringing high standards to technology has sped up our way of life. The largest effects that technology has had on men and women have come through the computer.

To say that this student struggles mightily with the English language is an understatement. No teacher in our program would rank the above two essays as equivalent, both scored 9—and thus passing—by WritePlacer *Plus.* The first student belongs in College English; the second belongs in our intensive, six-credit version of College English. I don't believe that this student simply needed more time to proofread his essay. There are far too many problems—and problems that indicate major syntax and usage problems—to believe that this student will quickly adapt himself to the relatively fast pace of College English. He will need the kind of one-on-one help that is both available and required in the intensive course. He needs someone who can help him understand the assumptions that he makes about how to communicate in writing.

If you don't buy my argument that the two essays above are of markedly different quality, there is no point in reading further. The chief reader at ACCUPLACER did not, for example; he concurred with the machine (Rickert 2002). However, if you do buy my argument, then the remainder of this chapter will focus on teasing out the kinds of problems that WritePlacer *Plus* seems blind to and the kinds of criteria it values instead.

It may be surprising that I must, at this juncture, admit that I believe WritePlacer *Plus* is a generally reliable placer of student essays. I am prepared to believe, as Vantage Learning (2000) asserts, that "the results [of their study] confirm earlier findings that IntelliMetric scores written responses to essay-type questions at levels consistent with industry standards and traditional expert scoring" (3). Timothy Z. Keith (2003), of the University of Texas at Austin, examines the validity studies of several automated essay-scoring systems and states that "IntelliMetric indeed produces valid estimates of writing skill" (158). (I would question Keith's use of the word "valid" here, but I will agree to the extent that reliability is one component of validity.) On 138 more or less randomly selected[3] essays, the average score assigned by my two readers and me agreed exactly with WritePlacer *Plus*'s score in 103 cases, agreed within one point in 33 cases, and agreed within two points in 2 cases. This is well within the reliability figure of 97 percent to 99 percent.[4]

How can I claim that IntelliMetric is both reliable and invalid at the same time? The focus of the remainder of this essay will be to point out the problems in validity that will cause some reliability problems *only when certain types of writing errors or problems occur.* Two of these errors were forecast by the paired essays above: the exaggerated value placed upon sheer length and the undervaluing of problems that have to do with readability.

THE PLACE OF LENGTH IN WRITEPLACER *PLUS'S* SCORING

As I mentioned earlier, we at Seton Hall University had developed the hunch that length seemed to be a disproportionally large factor in scoring. Faculty evaluating WritePlacer *Plus* at Middlesex County College have developed a similar intuition, that ACCUPLACER overvalues length in scoring student essays (Lugo 2005). To test such a hypothesis, a standard statistical method called regression analysis can easily be applied. Regression analysis calculates the variance in the WritePlacer *Plus* score that is accounted for by length, in this case, number of words per essay. An analysis of 221 randomly selected essays from 2002 through 2004 showed that fully 85 percent of the variance in essay scores was due to length.[5] This is a figure far higher than my intuition had led me to believe, and the implications are substantial. First, it means that length is valued far more than any teacher of writing would value it. It's true that, at the level of a first draft, length is an important indicator of fluency. In a way, it's a pleasant surprise to see that a testing company would value sheer length, as opposed to correctness, since length is related to fluency and idea development. However, 85 percent seems too high.[6]

The hypothesis that WritePlacer *Plus* valued length over all other variables was confirmed when I appended one essay to another—simply copying and pasting two essays together and submitting the combination as a single essay—to see how the score would change. In the first case, I appended two essays that each scored a 7, but each component essay had a position that contradicted the other. The first argued that taking two jobs or working overtime was a fine choice to make, while the second argued that making such a decision would ultimately be destructive. The result? An essay that scored a 10. In the second case, I appended two essays that also scored 7s, but in this case the two essays were on two entirely different topics, one on the most significant technology and the other on the advisability of working two jobs. The result? An essay that scored a 9, including a 9 on the focus subscore.

Of course, students don't naturally append essays of opposite points of view or of different topics altogether, but they do have problems recognizing when they have contradicted themselves and when they have gone off topic. These experiments provide some indication of how unlikely WritePlacer *Plus* is to "notice" the difference between essays that are well focused and essays that aren't. Or, at the least, WritePlacer *Plus* will value length so greatly that differences in focus may not show up even when they're egregious. It is hard to imagine a human reader so taken by the sheer verbiage in a piece of writing that he or she wasn't far more put off than was the computer by a complete and inexplicable switch in point of view or topic. Focus is one of five subcriteria upon which WritePlacer *Plus* scores the essays, yet my experiments suggest that focus takes a distant second place to the criterion of length.

Another experiment shows that WritePlacer *Plus* cannot judge the difference between concise and bloated language. An essay can be more confusing because of redundancy and superfluity and still score higher, because of length, than a concisely written essay. The first of the following two excerpts is the original introductory paragraph from a concisely written essay that scored a 6 (176 words). The second excerpt is the same paragraph that I loaded with bloat; the entire essay, filled throughout with such verbiage to reach 276 words, scored an 8.

> Many technological changes have occured since the formation of this country. The invention of the automobile has had a larger effect on the United States than any other invention.

> Many technological changes have occured since the formation of this country. Lots of changes have happened ever since our country first started. The invention of the automobile has had a larger effect on the United States than any other invention, even though there are indeed lots of inventions worth talking about.

One of the attributes of writing that English teachers prize is clarity. Generally, this means weeding out the extraneous words and phrases that do not contribute directly and powerfully to the idea at hand. If 85 percent of what WritePlacer *Plus* values is length, it's impossible for it to value concision as well. In the experiment above, I carefully padded the sentences to add absolutely nothing useful to the original phrasing, often merely rephrasing a sentence to create pure redundancy. The inability to detect the difference between spare and bloated writing explains why Nancy and I both passed some essays that received a 7, a failing score, from WritePlacer *Plus.*

The message for any high school seniors reading this essay is clear: write more and you'll pass. Specifically, write at least 400 words, if your institution has a cutoff of 8, to be placed in College English. Of the 208 essays that I examined myself, no essay of more than 373 words received less than an 8. This is hardly a large number of words, considering that the directions for writing the essay stipulate that the essay should be between 300 and 600 words.

CORRECTNESS

Herrington and Moran (2001) suspected that computers evaluated essays for length and correctness. They were certainly right about length. I believe they are partially right about correctness.

The following is the first paragraph from Andy's essay, perhaps the most error-filled non-ESL essay of the batch we reviewed.

> With all the different types of technological advancements that has changed the face of the world as we no it, the most influencial thing to be the cumputer. the reason that i say the cumputer is because the cumputer has the ability to organize, meanig file, alphabitized, and even manage. the cumputer in my opinion is the saving grace of the twenty first century.

Andy's essay received a 6 overall, with 5s for both the sentence structure and conventions subscores. I edited his essay to eliminate all the mechanical and grammatical problems (though not other, less obvious, word-choice problems, like "influential thing"), yielding a first paragraph that looks like this:

> With all the different types of technological advancements that have changed the face of the world as we know it, the most influential thing has to be the computer. The reason that I say the computer is that the computer has the ability to organize, meaning file, alphabetize, and even manage. The computer, in my opinion, is the saving grace of the twenty-first century.

The revised essay received an 8 overall, with 8s for the sentence-level subscores. This dramatic improvement suggests that WritePlacer *Plus* does indeed pay attention to correctness.

However, correctness is fairly narrowly conceived in WritePlacer *Plus.* The changes in spelling, subject-verb agreement, punctuation, sentence structure, and capitalization do make a difference in how this essay reads. But all these changes do not make as much difference as the editing to the anomalous essay cited in full at the beginning of this chapter. In its original form, Carl's essay received a 9, with subscores of 9 for sentence structure and 8 for conventions—a passing score at the vast majority of colleges that use ACCUPLACER. My revision, which involved drastic editing, resulted only in a 10, with subscores of 10 for both sentence structure and conventions. The readability problems for Carl's essay are at least as great as for Andy's—and require far more substantive editing. I changed 52 percent of the words from the original in Carl's essay, in contrast with only 21 percent of the original in Andy's essay. Yet the score went up only one point; WritePlacer *Plus* appears to have had a harder time "noticing" errors in Carl's essay. This may be due to the type of error. Of the 65 words I changed in Andy's essay, 40 were spelling or capitalization errors.

To learn whether WritePlacer *Plus* has problems "noticing" the types of errors in Carl's essay, I edited it only for the relatively few spelling and mechanical errors it has. When I finished editing, Word showed no green or red underlining, hence no spelling errors and no obvious grammatical errors, and yet a great number of syntactic and word-choice problems remained. Nevertheless, the new *limited* revision scored a 10, albeit with slightly lower subscores (10, 8, 8, 9, 9 vs. 10, 9, 8, 10, 10). The new revision had all the old readability problems but scored just as high as the far more thorough revision I had submitted earlier. WritePlacer *Plus* appears to value the combination of length plus mechanical correctness over concision and clarity.

Unless an essay has sentence-level problems of a certain type—generally mechanical—WritePlacer *Plus* has a hard time noticing them. Carl's essay required so much revision that only 231 words of the original 486 (or 48 percent) remained. By comparison, Andy's essay included 240 words of the original 305 (or 79 percent). The difference was that the 65 words that were changed were the kind that WritePlacer *Plus* "noticed."

Analysis of another essay confirms this finding. Gail's essay has a different sort of error—not as mechanical as in Andy's essay and not as subtle and pervasive as in Carl's essay. This excerpt reveals the kind of error this student is prone to:

> First of all when you apply for applications, jobs ask you to list your computers skills. Reason being most office work deals with online communications or some other type of knowledge of software. So if you have no experience in the field than you won't be able to get the job done. Then this person is left mad and disappointed because when he/she was growing up computers weren't used. Now that can be hard on a person who needs work to provide for their family. Computers being fatal to the workforce can be hard for a non computer literate person to get a job.

Table 1 (next page) indicates the types and number of errors in her essay. (The labels may be disputed in some cases, and the significance of Gail's use of "you" may be debated, but the overall number and significance should not be in question.)

I edited out 46 errors. The result? Absolutely no change in score, not even in the subscores, despite changing 114 of 398 words in the original essay (29 percent), more than I changed in Andy's essay. Gail's essay received a 9 overall, with subscores of 9, 8, 8, 9, and 9. Her errors, perhaps not coincidentally, do not all appear in Word. There are only four green underlined phrases, and only one in the excerpted passage above. Perhaps WritePlacer *Plus* has as much difficulty noticing these types of problems as Word does. In the above passage, no one would

TABLE 1
Errors in Gail's essay

Type of error	*Number of errors*
Subject/verb agreement	1
Fragment	4
Inappropriate informality (you)	8
Spelling/typo	7
Word choice	9
Wordy language	7
Switching persons	2
Faulty antecedent	2
Sentence structure	1
Comma splice	2
Verb form	1
Missing word	1
Misplaced modifier	1
Total	46

question the logic problem in the sentence "jobs ask you to list your computers skills." And the change between "you" and "this person" represents a confusing switch in person. The great majority of errors in Andy's essay that I changed were purely mechanical—mostly spelling—yet changing those relatively few errors had far more impact on the score than changing the wide variety of largely nonmechanical errors in Gail's essay.

In conclusion, WritePlacer *Plus* "values" correctness narrowly conceived. It certainly picks up spelling and other obvious grammatical problems, but it does not pick up more subtle differences having to do with word choice and syntax, differences that often make a greater difference in the readability of an essay. Especially when the values of concision and length are at odds—and they often are—WritePlacer *Plus* does not make meaningful distinctions.

Finally, the astute reader may notice that I have used no ESL essays to make my argument in this section. Some of the errors may sound like ESL errors—and it's true that Gail is an African-heritage student, possibly explaining why her writing has certain errors, like subject-verb agreement, in common with ESL writers—but the problems are not confined to ESL errors. Nevertheless, the most egregious scoring of essays occurs with ESL student essays. Here is an excerpt from Juan's essay, which received an 8:

> Technology across the countries,is and will the fist thing to discovery more important idead and permit to grow in many aspecs of the humanlife The general purpose technologies is to have much succesful improvement and eventually comes to be used for many people aroud the word, to have many uses, and to have many technological complementarities. The most cited examples include electricity, computers tv, E.T.C. Second, because thechnological changes give to as big information to undestand was happened in the word, thought the media, but also our technology can destroy many beuty things include human life.

On the one hand, I cringe to place an excerpt from Juan's essay here because his writing seems more thoughtful and engaged than many students', but on the other hand, he would not have been well served in a College English I classroom.

SENTENCE-SKILLS SCORE AS CORRECTIVE?

Although WritePlacer *Plus* clearly does not take into account the full extent of sentence problems in scoring placement essays, its sentence-skills test often provides useful information to alert the administrator to potential writing problems. Of twenty anomalous essays in which my readers and I all agreed that the frequency of errors made the WritePlacer *Plus* score unlikely, in thirteen cases the writers had sentence-skills scores that were more than 10 points below the average for their essay score. In three cases the sentences-skills scores were 7 or 8 points below. In only four cases were the sentences-skills scores the same or higher than we would have expected. In Table 2 (next page), note that the sentences-skills score (out of 120) is usually far below the average sentence-skills score for a given essay score. Notice also that in a few cases, the essay score is anomalously *higher*, and in these cases the sentence-skills score is significantly higher as well.

The differences between essay score and sentence-skill score are especially pronounced for those students who have English as a second language. Remember the excerpt from Juan's essay (essay F in table 2), for example, which received an 8—a passing score at Seton Hall—and compare it to this paragraph from Kim's essay (essay M in table 2), a more normal-quality 8:

> These days, without a proper education and some luck, it is beyond impossible to receive a decent job. People need to work hard to support not only themselves but their families. That may mean getting a second job

TABLE 2

Anomalous essays with associated sentence-skill (SS) scores

Essay	WritePlacer Plus score	Readers score	Sentence-skills score	Average SS score for a given essay score	Sentence-structure subscore	Conventions subscore	ESL status
A (Gail)	9	7	86	97	9	9	No
B (Carl)	9	7	75	97	9	8	No
C	7	6	91	88	6	6	No*
D	7	6	80	88	7	6	No
E	5	4	29	67	5	4	Yes
F (Juan)	8	5/6	35	94	7	7	Yes
G	11	8	64	103	11	10	Yes
H	9	7	33	97	8	8	Yes
I	12	9	66	103	11	10	Yes
J	6	7	104	85	6	6	No
K	6	7	101	85	5	5	No
L	8	7	35	94	8	8	No
M (Kim)	8	7	104	94	8	7	No
N	8	9	110	94	8	8	No
O	9	7/8	76	97	8	8	Yes
P	9	8	96	97	9	9	No
Q	9	7/8	119	97	9	9	No
R	11	9/10	68	103	10	10	No
S	12	11	96	103	12	11	No
T	5	4	47	67	5	6	Yes
U	11	10	96	103	10	10	No
V	11	10	83	103	11	10	No
W	12	11	92	103	11	10	No

*Language other than English spoken at home

> or working overtime. Although, working more then the average person may make one more tired and sleepy, it does not necessarily make them unhealthy or anti-social. If one can manage their time wisely, there is plenty of time in a day to work as much as needed and to find some time in between for your families and friends.

Fortunately, the sentence-skills score for each student helped clarify the placement. Juan received a 35; Kim received a 104.

Unfortunately, in many cases, a lower sentence-skills score does *not* accurately predict an anomalously lower-quality essay. And higher sentence-skills scores rarely predict anomalously higher-quality essays.

Understandably, when I added the sentence-skills score to the word count in the regression analysis, it accounted for only an additional 1 percent in the variance of the WritePlacer *Plus* grade. Still, the question remains, why would a computer-scoring system that includes sentence structure and conventions in its criteria appear to ignore the kinds of problems evinced by ESL students and even English speakers who have significant sentence-level problems? Note that the sentence-structure and conventions subscores for each of the students in the above table are either the same or just one lower than the overall score (with one exception).

ORDER AND COHERENCE

If WritePlacer *Plus* overvalues the most macro criterion, length, and the most micro criterion, mechanical correctness, how does it fare with the criteria that lie between these two: order and coherence? (See McGee's "Experiment 1" section in chapter 5 of this collection for a related analysis.) When we read the entire collection of essays, problems of organization and cohesiveness struck us less than complexity of thought, depth, and length. However, I was struck on numerous occasions by essays that seemed especially strong because they seemed more like an argument than a list of points.

Dan's essay was rated superior to the rest of the 8s we examined. Lengthwise, at 364 words, it ranked in the middle third of all the 8 essays. Its subscores were 8 for focus, sentence structure, and conventions; 7 for development and organization. Thus, it was rated an 8 but not a particularly strong 8. However, it had a quality that many other essays, even 9s, did not have: a sense of argument, a building of statements toward a conclusion, so that sentences and paragraphs had a uniquely possible position in the overall scheme of the essay. Sentences and paragraphs could not be randomly distributed and still make sense.

In contrast to Dan's essay, here is an example of a paragraph from Larry's essay, rated 9, that reads just as well (or not well) in a very different order. The original:

> (1) Basically everything in America is organized by computers. (2) Important places like banks rely heavily on computers for business. (3) Computers have contributed a great deal to criminal justice as well. (4) Computers are capable of storing massive amounts of information and scanning through information rapidly. (5) Solving crimes has been made easier because everyone's background and fingerprints can be saved and if a search of a person to a particular fingerprint is needed, it can be matched.

Here is a version with sentences in this order: 3-5-2-4-1.

> (3) Computers have contributed a great deal to criminal justice as well. (5) Solving crimes has been made easier because everyone's background and fingerprints can be saved and if a search of a person to a particular fingerprint is needed, it can be matched. (2) Important places like banks rely heavily on computers for business. (4) Computers are capable of storing massive amounts of information and scanning through information rapidly. (1) Basically everything in America is organized by computers.

In fact, one might argue that this order is an improvement, but the point is that there is nothing compelling about the order in the original version.

Dan's essay, on the other hand, suffers a good deal if rearranged. Look, for example, at his second paragraph:

> (1) Unfortunately, many people in the world have to work multiple jobs in order to survive. (2) As a result, their quality of life is harmed extremely. (3) Their relationship with their family suffers because they rarely see them. (4) Their marriage (if it lasts) suffers largely due to the lack of quality time that the person spends with their spouse. (5) As a result of their relationships suffering, the person may develop emotional or physical problems due to the stress. (6) All this happens just because of working too much.

The preceding paragraph in Dan's essay had set up the importance of achieving balance between work and relaxation. Thus the word "unfortunately" logically alerts us to the first sentence of the above paragraph. None of Dan's other paragraphs could logically go here. The second sentence, beginning with "As a result," does in fact logically follow "as a result" of the first proposition. Sentences 3 and 4 must follow sentence 2 because they both provide examples of how the quality of life can be harmed (even though sentences 3 and 4 themselves could be switched). Sentence 5 summarizes sentences 3 and 4 and identifies the consequence. Sentence 6 offers a conclusion that echoes the theme of the first sentence.

But Dan's essay received an 8, whereas Larry's essay, coming in at 298 words, received a 9. How could I test my hunch that sentence order doesn't "matter" to WritePlacer *Plus*? The solution seemed obvious: order the sentences randomly to see what effect there was on the scoring. I cut out pieces of paper, numbered them from 1 to 21, shuffled them, then created a new essay based on the new order. To give you an idea of how unsatisfying it would be to read this essay, I give you the first several sentences:

> Too much work usually results in stress, and stress is harmful to one's quality of life. I belive that too much work is only harmful to one's life. Unfortunately, many people in the world have to work multiple jobs in order to survive. As a result of free time, one has a chance at a good life and therefor are less likely to undergo physical and emotional stress. All this happens just because of working too much. As a result of their relationships suffering, the person may develop emotional or physical problems due to the stress.

It's still possible to discern a topic and even a point of view, but there is certainly no clearly demarcated progression of ideas with evidence, as I saw in the original. Yet this piece of writing received the same score, an 8. In fact the subscores were the same, except that the randomly ordered essay received a 7 instead of an 8 for conventions. The results of this experiment appear to challenge Vantage Learning's claim that IntelliMetric examines "transitional fluidity and relationships among parts of the response" (Vantage Learning 2004a, 3).

Not content with one example, I took another essay, a 7, that was composed of very short paragraphs (mostly one sentence each) that had a definite structure to them. It might be synopsized as follows: "Technology has greatly impacted us, making our lives easier and more enjoyable. For example, consider instances A, B, C. However, instance D is the most important for reasons X, Y, and Z. Thus the impact of the Internet has been large and positive. As for those who opposed technology, they can't stop it but they can adapt to it." Not a brilliant essay, but one in which order actually counts. I reordered the essay deliberately to undo the rhetorical effectiveness of the order. WritePlacer *Plus* scored the original essay a 7 (7, 6, 6, 6, 6). It scored the reordered paragraphs a 7 (7, 6, 6, 7, 6). Notice that the score for organization, the third criterion, remains the same.

If randomly ordering the sentences of an essay that is reasonably effectively ordered results in no change of score, it's hard to understand what WritePlacer *Plus* counts as "organization."

RETHINKING THE PLACE OF LENGTH AS A CRITERION

What are we to make of WritePlacer *Plus*'s apparent inability to discriminate among texts that seem so different? One possibility is that the problem lies not in the computer but in the human scorers. After all, ACCUPLACER's "chief reader" read Carl's essay and agreed with WritePlacer *Plus* that it is "a very good writing sample that substantially communicates a whole message to a specified audience. . . . The writer competently handles mechanical conventions such as sentence structure, usage, spelling and punctuation, though very minor errors in the use of conventions may be present" (Rickert 2002). However, no one to whom I have shown this essay at Seton Hall or elsewhere would place the writer in College English. Certainly no one, even ACCUPLACER's chief reader, would give Juan, the ESL writer, an 8 for his essay or think that College English I was an appropriate course for him. But this doesn't eliminate the possibility that ACCUPLACER's essay readers have been normed to value development of idea (length) over sentence-level readability.

For those involved in the development of computer scoring over the past thirty-five years, my finding that length is so prominent a factor may come as no surprise, since early on length was identified as the most reliable measure of the quality of holistically scored essays (Huot 1996; Powers et al. 2002; Roy 1993). However, current technology promises to move beyond the use of mere length to evaluate essays. Vantage Learning (2004e) recently put out a press release extolling the virtues of its product, implying that we have entered a new era in machine scoring: "IntelliMetric is the world's most accurate essay scoring engine, using a rich blend of artificial intelligence (AI) and the digitization of human expertise to accurately score and assess examinee responses to open-ended essay questions in a range of subjects." Yet in fact IntelliMetric appears to have little ability to discriminate between essays that are bloated or concise, ordered well or chaotically, focused on the same topic or on entirely different topics, written in clear prose or marred throughout by nonsimple errors. It is unfortunate that, because Vantage Learning wants to keep proprietary information secret, it will not share the information that would help educators understand why there would appear to be such a gap between Vantage Learning's claims and the findings in this chapter.

In the final analysis, the experiments above offer strong evidence that IntelliMetric cannot really "read" for the criteria that ACCUPLACER says it can. They suggest, instead, that it discriminates according to length and a limited number of mechanical and grammatical errors. What is remarkable is that, in so doing, WritePlacer *Plus* is able to reliably place the great majority of students. As much as I'm concerned with WritePlacer *Plus*'s ability to evaluate essays where superior length is not matched by superior coherence, correctness, or concision, the discovery that length is a far more reliable predictor of quality than anybody—even ACCUPLACER—would have expected forces us to rethink the relationship between fluency, the ability to get words down on paper, and quality in writing. How is it possible that 85 percent of the variance in WritePlacer *Plus*'s scores is attributable to the length of the essay? It suggests that sheer fluency—even in an *untimed* test—is extraordinarily important. Sheer fluency correlates, in the great majority of cases in which WritePlacer *Plus* is reliable, with focus, organization, coherence, sentence structure, and grammatical and mechanical correctness. The implications of this discovery and the connections to composition theory lie beyond the scope of this essay, but it would seem to confirm the priority that the "process approach" places on freewriting and journaling and the emphasis that WAC gives to informal writing as a vehicle for helping students make sense of constructs within the disciplines. That is, there is much value in doing lots of writing—not, of course, without a meaningful context, but writing to explore, to make sense of, to make connections, to play around with words.

IMPLICATIONS FOR PRACTITIONERS

In practical terms, the concerned test or writing program administrator will not be able to accurately place all students based on the essay score alone. The sentence-skills score will be of some real help in alerting the placement administrator to potential problems. So will questions about whether the student's first language is English and whether other languages are spoken at home. Since we saw occasional 11s that were questionable and occasional 7s that we considered passes (at least in conjunction with the reading and sentence skills), we had to do a fair amount of spot-checking to feel comfortable with all the placements. We also encouraged retesting if students believed the tests did not accurately represent their abilities. Of course, none of the discussion here takes up larger validity questions related to one-shot placement tests, as opposed to portfolios or directed self-placement, for example.

Increasingly, computer scoring is making its way into the K–12 educational market. At the 2005 New Jersey Writing Alliance conference, about twenty instructors from both college and high school attended an interest group on machine scoring of essays. They wanted to know whether computers might be useful in helping them ease their monstrous teaching loads in some way. Other reports confirm the interest of teachers and entire school districts using AES in the classroom (Borja 2003; Manzo 2003). MY Access! is an online instructional writing program, based upon IntelliMetric, that Vantage Learning has offered for a few years now at all levels from elementary through high school.

A Vantage publicist explains that MY Access! "provides immediate diagnostic feedback to engage and motivate students to write more and improve their composition skills" (Vantage Learning, 2005b). Scott Elliot, COO of Vantage Learning, claims, "Teachers can focus on specific strengths and weaknesses, by domain" using MY Access!'s "highly prescriptive feedback" (Vantage Learning, 2005b). The findings in this chapter certainly raise questions about IntelliMetric's ability to give meaningful diagnostic feedback. Its holistic scoring and assessment of certain mechanical and grammatical problems may be trusted, but the evidence here suggests that IntelliMetric is unable to make judgments about order, word choice, certain grammatical errors, and focus. Using MY Access! as a way to accurately assess even a modest range of problems in a student's text would appear unwise.

The analysis in this chapter suggests that, as many English teachers would like to believe, the computer cannot really read—or even simulate reading—but it also suggests how reliability and validity can be separated to some degree. If most of the quality of an essay is directly attributable to length and some mechanical and grammar errors, then the great majority of essays will be scored reliably. Only a relatively few essays—those that are loaded with lack of coherence, loss of focus, and the more subtle syntax and word-choice problems that IntelliMetric cannot "see"—will reveal the validity problems that are always present but usually hidden, masked by the computer's ability to do a few simple tasks consistently.

ACKNOWLEDGMENTS

I want to thank Gita DasBender and Christy Guerra, both writing instructors at Seton Hall University, for their enthusiastic participation in this research, as well as acknowledge the support of the Seton Hall University Research Council.