Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: A correlational study. *Contemporary Issues in Technology and Teacher Education*, *8*(4), 310-325.

Automated Essay Scoring Versus Human Scoring: A Correlational Study

Jinhao Wang
South Texas College

Michelle Stallone Brown
Texas A&M University-Kingsville

Abstract

The purpose of the current study was to analyze the relationship between automated essay scoring (AES) and human scoring in order to determine the validity and usefulness of AES for large-scale placement tests. Specifically, a correlational research design was used to examine the correlations between AES performance and human raters' performance. Spearman rank correlation coefficient tests were utilized for data analyses. Results from the data analyses showed no statistically significant correlation between the overall holistic scores assigned by the AES tool and the overall holistic scores assigned by faculty human raters or human raters who scored another standardized writing test. On the other hand, there was a significant correlation between scores assigned by two teams of human raters. A significant correlation was also present between AES and faculty human scoring in Dimension 4 - Sentence Structure, but no significant correlations existed in other dimensions. Findings from the current study do not corroborate previous findings on AES tools. Implications of these findings for English educators reveal that AES tools have limited capability at this point and that more reliable measures for assessment, like writing portfolios and conferencing, still need to be a part of the methods repertoire.

An increasing number of school districts and higher education institutions are adopting Automated Essay Scoring (AES) to assess students' writing for placement or accountability purposes (Shermis & Burstein, 2003b; Vantage Learning, 2001b). The Educational Testing Service (ETS) has used its AES tool "e-rater" to replace one of the two human graders for the writing portion of the Graduate Management Admission Test (GMAT) since 1999 (Herrington & Moran, 2001). The College Board and ACT testing companies have used Vantage Learning's AES tool IntelliMetric™ to rate the WritePlacer *Plus* test and the e-Write test, respectively (Haswell, 2004). The obvious advantages of using AES tools for large-scale assessment include timely feedback, low cost, and consistency of scoring. Additionally, if applied to classroom assessment, AES tools can reduce the workload of writing instructors and offer immediate feedback to every student (Bull, 1999).

However, whether AES tools can assess writing in the same way as do human raters is an issue that has been continuously debated. Testing agencies and AES developers have published numerous research results that generally show high agreement rates and strong correlations between AES scores and human raters' scores, yet the predictability rates have been low (Haswell, 2004). Because writing assessment is intimately related to teaching, learning, and thinking, the use of AES tools has caused much concern from composition scholars, who fear that the approaches taken by AES tools may send the wrong messages to students about the nature of writing.

For example, Herrington and Moran (2001) believed that writing had the power to change a person and the world, but if machine scoring were adopted, students might not be able to understand the power of writing because they would feel they were merely writing to the machines. The same concern was expressed by Drechsel (1999), who believed that students wrote with the expectation of getting human reaction. If they wrote to the machine, their "voice" would get lost—they would be "writing into silence" (p. 380).

As writing assessment entails evaluation of writing features valued by writing instructors, AES directly impacts writing instruction, and scholars fear an AES approach may change the main focus of the writing instruction, misleading instructors to focus on "discrete stylistic components" rather than focusing on writing within "communicative contexts" (Fitzgerald, 1994, p. 16). In other words, writing instructors may teach writing as if it is related to counting rather than meaning making (Cheville, 2004). Besides, it is nearly impossible for AES tools to imitate the human assessment process, which involves "multiple subjectivities" and "sophisticated intellectual operations" (Anson, 2003, p. 236).

The ongoing debate about the nature of AES and its implications on writing instruction and writing assessment necessitates more research in the validity and usefulness of AES tools. However, the realm of AES research has so far been occupied by commercial testing companies. It is important that potential users of AES in secondary and higher education begin to direct their attention to investigating how AES works and to what extent AES can replace human raters, since both writing instruction and students' learning are at stake.

To understand how AES tools work, it may be helpful to take a look at how AES tools evolved and how some of the major AES tools are currently functioning (a glossary of terms is provided in the appendix). In 1966, Ellis Page, the inventor of Project Essay Grader (PEG) and the pioneer of AES, published an article entitled "The Imminence of Grading Essays by Computer." In this article Page described his invention of using computer technology to grade essays and expressed his optimism about the promising future of relieving English teachers from the burden of grading papers (Wresch, 1993).

Page's PEG uses three steps to generate scores (Yang, Buckendahl, Juszkiewicz, & Bhola, 2002). First, it identifies a set of measurable features that are approximations or correlates of the intrinsic variables of writing quality (proxes); second, a statistical procedure—linear multiple regression—is used to find out the "optimal combination" of these proxes that can "best predict the ratings of human experts" (Yang et al., 2002, p. 394); third, the proxes and their optimal combination are then programmed into the computer to score new essays.

Other AES tools use similar three-step strategies to score essays. Intelligent Essay Assessor (IEA), which is used by the ETS to score the Graduate Equivalency Diploma essay test, grades essays by using the technique of latent semantic analysis—it first processes a large body of the texts in a given domain of knowledge, establishing a "semantic space" for this domain. Then, it analyzes a large amount of expert-scored essays to learn about the desirable or undesirable qualities of essays. Finally, it uses a factor-analytic model of word co-occurrences to find the similarity and semantic relatedness between the trained essays and the new essays at different score levels (Rudner & Gagne, 2001; Yang et al., 2002).

E-rater, which was also adopted by ETS, uses natural language processing and information retrieval to develop modules that capture features such as syntactic variety, topic content, and organization of ideas or rhetorical structures from a set of training essays prescored by expert raters. It then uses a stepwise linear regression model to find the best combinations of these features that predict expert raters' scores. These combinations are processed into the computer program to score new essays (Yang et al., 2002).

Building on the strategies utilized by PEG, IEA, and e-rater, IntelliMetric™, developed by Vantage Learning, incorporates the technologies of artificial intelligence and natural language processing, as well as statistical technologies. These combined approaches are treated as a "committee of judges," and "potential scores" from these judges are calculated by using proprietary algorithms to achieve the most accurate score possible (Vantage Learning, 2003, p. 9). Capable of analyzing more than 300 semantic, syntactic, and discourse level features, IntelliMetric functions by building an essay scoring model first—samples of essays with scores already assigned by human expert raters are processed into the machine, which would then extract features that distinguish essays at different score levels. Once the model is established, it is validated by another set of essays. Finally, it is used to score new essays (Elliot, 2003).

At the present stage, AES tool developers are still exploring ways to enhance the correlation between writing quality and surface features of writing, such as "lexical-grammatical errors," or "rough shifts," or "rhetorical relations" (Kukich, 2000, p. 26). However, technologies such as artificial intelligence and natural language processing need to become more sophisticated before AES tools can come closer to simulating human assessment of writing qualities. In terms of evaluating the content of essays and assessing works written in nontesting situations, AES tools are still lagging behind human raters (Warschauer & Ware, 2006).

The Purpose of the Study

The purpose of the current study was to analyze the relationship between AES and human scoring in order to determine the validity and usefulness of automated essay scoring for large-scale placement tests. Specifically, the researcher examined the validity of one automated essay scoring tool patented as IntelliMetric, which was used to score College Board's WritePlacer *Plus*—an online standardized writing test. A group of Developmental

Writing students were invited to take WritePlacer *Plus*, and their scores assigned by IntelliMetric were compared with their scores given by human raters on the same test. In addition, the study also examined the construct validity (whether the results of an instrument correlate well with the results of other instruments that measure the same construct) of IntelliMetric by comparing the performance of IntelliMetric with human scoring of written responses to another standardized writing test, the Texas Higher Education Assessment (THEA), taken by the same group of students. Results from the study might help institutions understand the implications of replacing human scoring with AES, so they can make informed decisions about which placement test or exit test to use.

Review of Literature

Since its advent in 1998, IntelliMetric has been adopted by many testing organizations, school districts, and higher education institutions. The College Board uses IntelliMetric to grade WritePlacer *Plus*, an online writing placement test. Other organizations such as Thompson Learning, Harcourt Companies, the states of Oregon and Pennsylvania, and the Secondary School Assessment Testing Board, have used IntelliMetric for various writing assessment needs (Vantage Learning, 2001b). Its popularity reflects the demand for the cost-effective and expedient means to evaluate writing tests as well as classroom writing assignments. However, the decisions made to adopt such an AES tool are still largely based on the research briefings published by Vantage Learning, the company that patented IntelliMetric.

The earliest research on IntelliMetric started in 1996 before IntelliMetric went into full operation. Since then, more than 120 studies have been conducted by Vantage Learning. The majority of them focused on verifying the validity of IntelliMetric by using various research designs, such as "Expert Comparison Studies," "True Score Studies," and "Construct Validity Studies" (Elliot, 2003, pp. 73-74). Most researchers have used the correlational study design to examine the holistic scores assigned by Intellimetric and human raters; a few researchers have examined analytic dimensional scores.

Nearly all the studies reported high correlation coefficient rates. For example, a study conducted in 2001 examined the validity of IntelliMetric in scoring essays written by entry-level college students, and it reported the Pearson *r* correlation coefficients for the six writing prompts as ranging from .50 to .83 (Vantage Learning, 2001a). Another correlational study of IntelliMetric scoring versus human scoring conducted in 2002 reported a Pearson *r* correlation coefficient at .77 (Vantage Learning, 2002a).

Several Vantage Learning studies focused on the correlations between IntelliMetric and human raters in both holistic scoring and dimensional scoring – analytic scoring on such features as Focus, Content, Organization, Style, and Convention. Sometimes, features varied according to the rubric developed by users, such as state testing agencies or school districts.

One of the dimensional studies examined how well IntelliMetric could be used in the Pennsylvania Student Skills Assessment Program. The results showed that for persuasive writing prompt, IntelliMetric had a slightly higher agreement rates (exact agreement rates) than did human raters in four dimensions—focus, content, organization, and style—whereas human raters had a higher agreement rate than IntelliMetric in one dimension—convention. IntelliMetric also had higher correlation rates with each of the human raters in the same four dimensions than human raters' correlation between themselves (Vantage Learning, 2000).

Vantage Learning researchers also published a study on dimensional scoring in 2002, and this time the focus was on validating IntelliMetric in grading WritePlacer ESL. The results showed that the overall holistic score of IntelliMetric had a strong correlation (.78 to .84) with human raters' scores. The exact agreement rates were moderately strong (52% to 58%). However, the dimensional scoring of IntelliMetric was "less reliable than holistic scoring," especially in the case of the "convention" dimension (Vantage Learning, 2002b, p. 4). In this dimension, the correlations between IntelliMetric scores and experts' scores across the prompts were .60 to .80, and the exact agreement rates (agreement between IntelliMetric scores and experts' scores) were 48% to 58%. For the remaining four dimensions, the correlation rates ranged from .72 to .89, whereas the exact agreement rates ranged from 44% to 60%. Both the highest correlation rate and the highest agreement rate came from the "content" dimension (Vantage Learning, 2002b, p. 5).

On the whole, all the Vantage Learning research reported strong correlations between IntelliMetric and human raters. However, these publicized studies were research briefs with no report on details in research design, such as how human raters were calibrated and whether significance tests were run to demonstrate the statistically significant difference between IntelliMetric's performance and human raters' performance. In addition, most of the studies were validation studies, with the validation data drawn from the same student population from which the training data were collected; thus, the generalizability of the findings from these studies was not demonstrated.

Research Questions

This study was guided by the following research questions:

1. How well do scores assigned by IntelliMetric correlate to scores given by human raters on the WritePlacer *Plus* test?

How well does IntelliMetric scoring correlate to human scoring in the overall rating of the essay?
How well does IntelliMetric scoring correlate to human scoring in measuring Focus?
How well does IntelliMetric scoring correlate to human scoring in measuring Development?
How well does IntelliMetric scoring correlate to human scoring in measuring Organization?
How well does IntelliMetric scoring correlate to human scoring in measuring Sentence Structure?
How well does IntelliMetric scoring correlate to human scoring in measuring Mechanics?

2. How well do scores assigned by IntelliMetric on the WritePlacer *Plus* test correlate to the respondents' scores on another standardized writing test, THEA, which is graded by human raters?

Methods and Procedures

Population and Sample

A sample of 107 students was randomly selected from an accessible population of 498 students enrolled in the highest level of developmental writing courses at a college in South Texas – a Hispanic serving institution with 95% of the students being Hispanic.

These students were required to take THEA as part of their developmental writing class to exit the program. The online writing test – WritePlacer *Plus* test – was given to these students as their practice test for the THEA writing exam.

Research Design

The current study utilized a quantitative correlational study design. Human scoring and automated essay scoring were selected as variables for computing correlation coefficients. Writing responses gathered from WritePlacer *Plus* test were graded by an automated essay scoring tool—IntelliMetric—as well as by trained human raters. Responses from the THEA writing test were graded by human raters trained by National Evaluation Systems (an automated essay scoring tool has not been adopted by the THEA program). For WritePlacer *Plus*, both the automated essay scoring tool and human raters assigned a holistic score for the overall quality of a writing response and analytic scores on five dimensions of each writing response—Focus, Development, Organization, Sentence Structure, and Mechanics. Human raters for both WritePlacer *Plus* and THEA assigned scores on a scale of 1 to 4, and the two raters' scores were added up to form a score of 2 to 8 for each essay. IntelliMetric assigned scores reflecting the sum of two raters' scores ranging from 2 to 8. Altogether, three sets of variables were examined in the correlational study, as indicated in Figure 1.

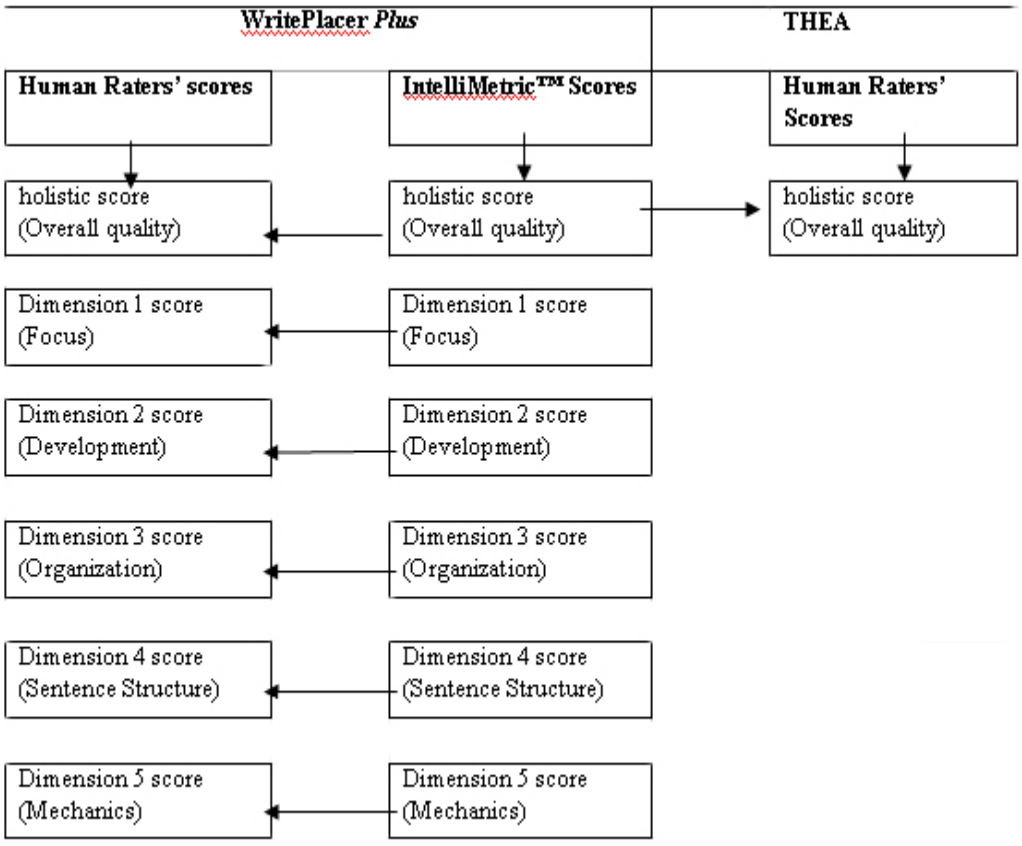

Figure 1. Correlational study design model.

Data Collection

To collect data, two standardized tests, WritePlacer *Plus* and THEA, were administered to the participants by trained proctors within a week's timeframe.

Participants took the WritePlacer *Plus* test first, and their writing samples were graded by the IntelliMetric instantly. These IntelliMetric scores were collected by the researcher and entered into the SPSS database. After the same group of students had taken THEA organized and proctored by the Testing Office at the college and after the THEA scores became available, the researcher obtained the score report. The participants' THEA scores were then entered in the SPSS database. At this point, the database was screened, and students who had a WritePlacer score only or THEA scores only were deleted. After the screening, 284 cases had both sets of scores and were kept in the SPSS database.

The SPSS Case Selection procedure was used to select a random sample of approximately 35% of 284 cases with the intention to get 100 cases, but this procedure yielded 107 cases. Then, the chosen 107 papers were assigned a number ranging from 1 to 107. These numbers were matched with students' names and college ID numbers, as well as their corresponding IntelliMetric scores and THEA scores. These 107 students' WritePlacer writing samples were then retrieved from the WritePlacer score report database. During the process of retrieving the writing samples, each paper was labeled with the assigned number and student's names and college ID numbers were kept anonymous, sealed with blank pieces of paper.

The retrieved writing samples were each graded by two trained human raters, who were volunteers from the Developmental English Department of the college where the research was conducted. To ensure consistency, both volunteers were instructors who had at least a master's degree in English or applied linguistics. They also had at least 5 years of experience in evaluating students' essays. Both volunteers had been to holistic scoring training sponsored by the National Evaluation Systems (NES). In addition, both of them received two more recent trainings.

The two raters graded each of the writing samples by first assigning a holistic score and then analytic score, to ensure that the holistic score was the rater's overall impression of the overall quality of the writing sample, not the average of the analytic dimension scores. After the grading was finished, the results were entered into the SPSS database with the other sets of results.

Data Analysis

First, the SPSS Explore procedure was run to examine the normality of the data. For the overall holistic scores assigned by the three scoring methods, namely, the IntelliMatric scoring of WritePlacer *Plus* (GRME1), the human rater (faculty) scoring of WritePlacer *Plus* (GRME2), and the human rater (NES experts) scoring of THEA writing test (GRME3), all 107 cases had valid scores, with no outliers. Table 1 displays the means, medians, and standard deviations for each scoring method.

Table 1
Descriptive Statistics for the Three Sets of Overall Holistic Scores, N = *107*

Variables	*M*	Median	*SD*
GRME1 (aeshs)	5.98	6.00	.87
GRME2 (hrhs_tot)	5.22	5.00	.965
GRME3 (theahs)	4.92	5.00	.963

The Spearman rank correlation coefficient test, a nonparametric version of the Pearson correlation coefficient test, was selected for determining the correlations between IntelliMetric scores and human raters' scores.

The Spearman rank correlation coefficient tests were run separately for analyzing the overall holistic scores and for each set of dimensional scores (SPSS 12.0 was used for data analysis). First, the bivariate correlation analysis was performed to evaluate the holistic score variables, namely, IntelliMetric holistic scores (aeshs), human raters' holistic scores on WritePlacer (hrhs_tot), and THEA holistic scores (theahs). The significance level was set at .05 divided by the number of correlations, using the Bonferroni approach to control for Type I errors.

Secondly, five Spearman correlation tests were run separately to analyze the dimensional scores given by IntelliMetric™ and faculty human raters on WritePlacer *Plus*. The significance level was set at .05 for each significance test.

Results

The results of the correlational analyses indicated that there was a statistically significant correlation between WritePlacer scores assigned by the faculty human raters and the THEA scores assigned by NES human expert raters (rs = .35, p < .017). On the other hand, the correlations between IntelliMetric assigned overall holistic scores and human raters' (faculty) overall holistic scores or THEA overall holistic scores were lower and not statistically significant. The detailed results are presented in Table 2.

Table 2
Correlations Among the Three Sets of Overall Holistic Scores, N = *107*

Variables	aeshs	hrhs_tot	theahs
aeshs	--		
hrhs_tot	.11	--	
theahs	.04	.35*	--
*p < .017			

Analyses of the dimensional scores showed that IntelliMetric scoring did not correlate well with human scoring in Dimension 1 - Focus, Dimension 2 - Development, Dimension 3 - Organization, and Dimension 5 -Mechanics, but it appeared to have a statistically significant correlation with human scoring in Dimension 4 - Sentence Structure (rs = 21, p < .05). The detailed results are displayed in Table 3.

Table 3
Correlations Between Dimensional Scores, N = *107*

Variables	*r*
D1aes – D1hr_tot	.16
D2aes – D2hr_tot	.17
D3aes – D3hr_tot	.06
D4aes – D4hr_tot	.21*
D5aes – D5hr_tot	.07
*$p < .05$	

Discussion

Results based on the correlational data analyses showed no statistically significant correlation between IntelliMetric scoring and human scoring in terms of overall holistic scores. This finding does not corroborate previous studies conducted by Vantage Learning, which reported strong correlations between IntelliMetric scoring and human scoring for overall ratings (Elliot, 2003). Nor does this finding support studies by some independent users published by Vantage Learning, such as Greer's (2002) study and Nivens-Bower's (2002) study, both of which reported strong significant correlations between IntelliMetric and human scoring. The different results produced by the current study seem to indicate that the IntelliMetric scoring model built by a pool of essays written by a different student population may not be generalizable to the student population in South Texas. The lack of significant correlation also raises the question whether IntelliMetric scoring can be consistent with human scoring at all times and in all situations.

On the other hand, the faculty human raters' scoring had a significant correlation with NES human raters' scoring of the THEA writing test (*rs* = .35), whereas IntelliMetric scoring did not show a significant correlation with the same NES human raters. This contrast seemed to indicate that the faculty raters performed more consistently with NES human raters than with IntelliMetric.

In terms of correlations between IntelliMetric scoring and human scoring in different dimensions of the essays, data analyses showed no statistically significant correlations in Dimension 1, 2, 3, and 5. However, a statistically significant correlation was found in Dimension 4 - Sentence Structure. These findings suggest that IntelliMetric seems to be more consistent with human scoring in assessing sentence structures, but not in assessing other dimensions. Again, these findings do not support the dimensional studies conducted by Vantage Learning, which reported strong correlations in focus, content, organization, and style for the persuasive writing prompt in one of the studies (Vantage Learning, 2000) and demonstrated the strongest correlation in the "content" dimension in another study (Vantage Learning, 2002b).

In general, findings from the study challenged the research results published by Vantage Learning, which demonstrated strong correlations between AES and human scoring. These findings also raised the question whether AES models built by writing samples from one student population were generalizable to writing samples from other student populations. If AES models are not generalizable, pending the confirmation of future

studies, then it may be necessary for a specific AES model to be built for a specific student population. In that case, the cost of using AES tools may become a question of concern.

Furthermore, the results of the current study pointed out the possibility of AES being significantly correlated to human raters in assessing Sentence Structure rather than in content-related features. If this finding is true, pending the confirmation of further studies, then it may mean that AES tools can be utilized more specifically in assisting student writers with feedback on improving their sentence skills.

Finally, the serendipitous finding from the current study indicating significant correlation between the two teams of human raters may mean human raters are more consistent with each other in assigning essay scores than with AES tools. A finding of this nature, if confirmed by future studies, may also call into question the validity of AES tools.

Implications

The correlational analyses, using the nonparametric test Spearman Rank Correlation Coefficient, showed that the overall holistic scores assigned by IntelliMetric had no significant correlation with the overall holistic scores assigned by faculty human raters, nor did it bear a significant correlation with the overall scores assigned by NES human raters. On the other hand, there was a statistically significant correlation, with an effect size of medium coefficient, between the two sets of overall holistic scores assigned by the two teams of human raters. Spearman Rank Correlation analyses of dimensional scores showed a significant correlation between IntelliMetric scoring and faculty human scoring in Dimension 4 - Sentence Structure but no significant correlations in other dimensions.

On the whole, the results from the current study support the conclusion that IntelliMetric did not seem to correlate well with human raters in scoring essays and that findings published by Vantage Learning did not appear to be generalizable to the student population in South Texas. The discrepancies between the findings of the current study and those published by Vantage Learning may be attributed to the following factors:

1. As the review of literature uncovers, most studies conducted by Vantage Learning were validation studies, which had split the pool of student writing samples from the same student population into two parts, using one part to build the scoring model and the other part as a validation data set. This means the scoring model built by the writing samples drawn from the same student population as the validation set might have extracted writing features idiosyncratic to the particular student population. Therefore, the scoring model could score the validation set with relatively high accuracy, but it is questionable whether its application is generalizable to other student populations who receive different writing instruction and have different writing experiences.
2. As the AES experts such as Kukich (2000) and Shermis and Burstein (2003a) acknowledged, artificial intelligence and natural language processing techniques utilized by AES have not yet reached the stage of perfection in simulating human intelligence and information processing, although the maturity level of AES techniques have been improving rapidly over the years. In the current study, IntelliMetric scoring yielded only a significant correlation with human scoring in Dimension 4 - Sentence Structure, so it is possible that automated essay scoring tools tend to be more accurate in evaluating surface features of writing samples at the sentence level. In contrast, human raters are trained to regard surface features as one of the five dimensions of writing; they may not have weighted surface features as heavily as did AES tools. More importantly, human raters emphasize meaning-making and communicative contexts (Cheville, 2004;

Herrington & Moran, 2001), which AES tools may still be incapable of identifying and evaluating.

As the interest in adopting AES tools increases, and as the development of AES technologies undergoes rapid changes, they still hold a promising future for writing assessment programs; therefore, continuous research and investigation in the validity and generalizability of the AES tools are inevitable. Based on the findings of the current study, further studies should be conducted to determine the validity and generalizability of the AES tools. Topics should include experimental studies that investigate which surface features impact the AES tools' assigning high scores, correlational studies that compare AES scores with multiple human raters' scores, correlational studies that compare participants' AES scores with their course grades, and comparative studies that examine the mean score differences across AES mean scores and two teams of human raters' mean scores, all on the same student writing samples. Qualitative studies should also be conducted to analyze essays that receive AES scores with a 2-point discrepancy from human raters' scores.

In the interim, school administrators who make decisions about what assessment tools to use need to take the validity of AES tools into consideration. While AES tools are cost effective and fast in returning results, they may not be as accurate as human raters in assessing students' written works. Therefore, it is a matter of choosing between efficiency and quality of assessment methods. Responsible decisions on the assessment tools should be based on multiple measures, which include not only timed writing samples assessed by human raters, but also students' writing portfolios and advising/counseling processes. In addition, for whatever assessment approach a school decides to adopt, a process for validating and evaluating the approach should be implemented to ensure that the assessment programs undergo continuous improvement for the sake of students' learning and success in their coursework (McLeod, Horn, & Haswell, 2005; Morante, 2005). The validation and evaluation methods should include studies that correlate students' placement scores with their course grade to examine the criterion-related validity, as well as interviews and surveys about students' and instructors' perceptions of the accuracy of the assessment.

For English teachers, AES tools have the potential to offer immediate feedback to students' writing while relieving the heavy load of grading. However, when the validity of the AES tools is still in question, the use of machine grading should be restricted to spelling checks and sentence skills feedback. Besides, if English teachers value writing as a communicative act, and if they truly want to improve students' writing, they will still need to assess students' writing personally and offer dialogic feedback to students, who will benefit not only from their teachers' specific comments but also from the human touch. The results here have important implications for English teacher education as well. While English educators may want to expose pre- and in-service teachers to AES tools, their utility is limited at this point, and well-documented assessment strategies, like writing portfolios and writing conferences, and a keen awareness of the process writing approach should still be included in English education methods courses and in the methods repertoire of practicing English language arts teachers.

Although an overwhelming grading load is often a reality for writing instructors, scholars such as Zinn (1998) have explored ways to ease the load. Zinn suggested using student-generated grading criteria and focusing on a couple of special grading problems. Instructors should also make writing assignment topics clear, so that the end product will be easy to grade. Sample papers and specific grading criteria will also assist with the grading process. Group responses and feedback to early drafts can also be used to help lighten the load. Too much commentary should be avoided.

Putting the issue of heavy grading load aside, the key issue behind writing assessment should continue to be the promotion of student learning and the improvement of the quality of students' writing. The means to achieve this end lies in the hands of human raters, rather than machines.

References

Anson, C.M. (2003). Responding to and assessing student writing: The uses and limits of technology. In P. Takayoshi & B. Huot (Eds.), *Teaching writing with computers: An introduction* (pp. 234-245). New York: Houghton Mifflin Company.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Harlow, England: Addison-Wesley.

Bull, J. (1999). Computer-assisted assessment: Impact on higher education institutions. *Educational Technology & Society, 2(3).* Retrieved from http://www.ifets.info/journals/2_3/joanna_bull.pdf

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *English Journal, 93*(4), 47-52.

College Board. (2004). *ACCUPLACER coordinator's guide*. New York: Author.

Cooper, C.R. (1977). Holistic evaluation of writing. In C.R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-31). Urbana, IL: National Council of Teachers of English.

Drechsel, J. (1999). Writing into silence: Losing voice with writing assessment technology. *Teaching English in the Two-Year College*, *26*, 380-387.

Elliot, S. (2003). Intellimetric™: From here to validity. In M.D. Shermis & J.C. Burstein (Eds.), *Automatic essay scoring: A cross-disciplinary perspective* (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum Associates.

Fitzgerald, K.R. (1994). Computerized scoring? A question of theory and practice. *Journal of Basic Writing, 13*(2), 3-17.

Haswell, R.H. (2004). *Post-secondary entry writing placement : A brief synopsis of research.* Retrieved from http://comppile.tamucc.edu/writingplacementresearch.htm

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63*(4), 480-499.

Kukich, K. (2000). Beyond automated essay scoring. *IEEE Intelligent Systems,15*(5),22-27.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. C. Burstein (Eds.),

Automatic essay scoring: A cross-disciplinary perspective (pp. 87-112). Mahwah, NJ: Lawrence Erlbaum Associates.

McLeod, S., Horn, H., & Haswell, R. H. (2005). Accelerated classes and the writers at the bottom: A local assessment story. *College Composition and Communication, 56 (4),* 556-579.

Morante, E. A. (2005, November). *Basic skills assessment and placement* [Powerpoint slides]. Presentation to the consulting meeting with South Texas College Assessment and Matriculation Taskforce, McAllen, TX.

National Evaluation Systems. (2005). *THEA faculty manual: A guide to THEA test results*. Retrieved from http://www.thea.nesinc.com/PDFs/THEA_FacultyManual.pdf

Nivens-Bower, C. (2002). *Establishing WritePlacer validity: A summary of studies* (RB-781). Yardley, PA: Author.

Rudner, L., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *ERIC Digest*, ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED458290).

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Page, E. B. (1966). The imminence of grading essays by computers. *Phi Delta Kappan, 47,* 238-243.

Shermis, M.D., & Burstein, J.C. (2003a). Introduction. In M.D. Shermis & J.C. Burstein (Eds.), *Automatic essay scoring: A cross-disciplinary perspective* (pp. xiii-xvi). Mahwah, NJ: Lawrence Erlbaum Associates.

Shermis, M.D., & Burstein, J.C. (2003b). Preface. In M.D. Shermis & J.C. Burstein (Eds.), *Automatic essay scoring: A cross-disciplinary perspective* (pp. xi-xii). Mahwah, NJ: Lawrence Erlbaum Associates.

Stutz, J., & Cheeseman, P. (1994, June). *A short exposition on Bayesian inference and probability*. Retrieved from the National Aeronautic and Space Administration Ames Research Center Web site: http://ic.arc.nasa.gov/ic/projects/bayes-group/html/bayes-theorem-long.html

Vantage Learning (2000). *A study of IntelliMetric™ accuracy for dimensionalscoring of K-12 student writing*. (Report No. RB-393). Newtown, PA: Vantage Learning.

Vantage Learning (2001a). *Applying IntelliMetric™ to the scoring of entry-levelcollege student essays.* (Report No. RB-539). Newtown, PA: Vantage Learning.

Vantage Learning (2001b). *IntelliMetric™ : From here to validity*. (Report No. RB-504). Newtown, PA: Vantage Learning.

Vantage Learning (2002a). *A study of expert scoring, standard human scoring and IntelliMetric™ scoring accuracy for statewide eighth grade writing responses.* (Report No. RB-726). Newtown, PA: Vantage Learning.

Vantage Learning (2002b). *IntelliMetric™ scoring for WritePlacer ESL* (Report No. RB-734). Newtown, PA: Vantage Learning.

Vantage Learning. (2003). *How does IntelliMetric™ score essay responses*? (Report No. RB-929). Newtown, PA: Vantage Learning.

Warschauer, M. & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*, 2, 1–24. Retrieved from the University of California - Irvine Department of Education Web site: http://www.gse.uci.edu/person/markw/awe.pdf

Wolcott, W., & Legg, S.M. (1998). *An overview of writing assessment: Theory, research, and practice*. Urbana, IL: National Council of Teachers of English.

Wresch, W. (1993). The imminence of grading essays by computers—25 years later. *Computers and Composition 10*(2), 45-58.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer automated scoring. *Applied Measurement in Education, 15*(4), 391-412.

Zinn, A. (1998). Ideas in practice: Assessing writing in the developmental classroom. *Journal of Developmental Education, 22*(2),28-34.

Author Note:

Michelle Stallone Brown
Texas A&M University-Kingsville
email: kfmns00@tamuk.edu

Jinhao Wang
South Texas College
email: jwang@southtexascollege.edu

Appendix
Definition of Terms

For the purpose of the current study, some technical terms are used and defined as follows:

Artificial intelligence – refers to the automation of activities normally associated with human thinking. Examples of such activities include decision-making, problem-solving, and learning. Artificial intelligence is, thus, the study of the computations concerned with intelligent behaviors. It is the study of how to make computers do things that human beings are still better at doing (Russell & Norvig, 2003).

Automated essay scoring – refers to "the ability of computer technology to evaluate and score written prose" (Shermis & Burstein, 2003a, p. xiii).

Bayesian analysis – refers to a mathematical method that calculates the probability of events by introducing prior knowledge into the calculations. This method was first developed by Thomas Bayes, an 18th century mathematician and theologian, who first published his Bayes' Theorem in 1763 (Stutz & Cheeseman, 1994).

Holistic scoring – refers to a grading method that utilizes a holistic scoring guide to rank order a piece of writing according to its overall quality or certain features. The rank ordering takes place "quickly, impressionistically, after the rater has practiced the procedure of rank ordering with other raters" (Cooper, 1977, p. 3).

Information retrieval – refers to the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information the user is interested in retrieving (Baeza-Yates & Ribeiro-Neto, 1999).

Latent semantic analysis – refers to a machine learning method that represents the meaning of words and passages through the use of statistical computations based on a large amount of texts. The underlying idea of this method is that a passage is "the sum of the meanings of its words" and that the combination of all the contexts in which a given word is present or absent determines the similarities of word meanings. Utilizing such concepts, Latent Semantic Analysis simulates human judgments and behavior in assessing the quality of the semantic content of an essay and analyzing essays "for the components of content that are and are not well covered" (Landauer, Laham, & Foltz, 2003, p. 88).

Natural language processing – refers to an interdisciplinary study of modern linguistics and artificial languages. It aims at building automated systems to understand natural language through automated syntactic and semantic analysis (Russell & Norvig, 2003).

Proxes – refer to the computer extractable predictive features that "approximate" the intrinsic features in an essay valued by human raters. The term was coined by Page (1966) when he designed the automated essay scoring system. A prox is a "computer-identifiable trait" that may "correlate with" the intrinsic value in an essay (Wresch, 1993, p. 46).

Texas Higher Education Assessment (THEA) – a test assessing students in three subject areas: reading, writing, and math. The writing portion of the test consists of two subsections. One subsection assesses students' ability to recognize elements of effective writing in a multiple-choice format. The other subsection assesses students writing ability in a multiple-paragraph essay format. The writing samples are graded by two human raters on a scale of 1 to 4. The sum of the two raters' scores constitutes a score ranging from 2 to 8 for each writing sample (National Evaluation Systems, 2005).

Trins – refer to the intrinsic features in an essay valued by human raters. The term was coined by Page (1966) when designing an automated essay scoring system to evaluate essays. Page believed the intrinsic features in an essay were extractable by computer programs and could be predicted by using statistical techniques. "A trin might be a human measure of values such as aptness of word choice" (Wresch, 1993, p. 46).

WritePlacer Plus— refers to a standardized writing test that measures writing skills at the level that is expected of entry-level college students. It is offered through the College Board's ACCUPLACER Program, and it is mainly an online writing test, but when requested, the paper-and-pencil version is also available (College Board, 2004).