J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 6, Number 2 · October 2007

Automated Essay Scoring Versus Human Scoring: A Comparative Study

Jinhao Wang & Michelle Stallone Brown

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 6, Number 2

Automated Essay Scoring Versus Human Scoring: A Comparative Study

Jinhao Wang & Michelle Stallone Brown

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Jennifer Higgins
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Wang, J. & Brown, M.S. (2007). Automated Essay Scoring Versus Human Scoring: A Comparative Study. *Journal of Technology, Learning, and Assessment, 6*(2). Retrieved [date] from http://www.jtla.org.

Abstract:

The current research was conducted to investigate the validity of automated essay scoring (AES) by comparing group mean scores assigned by an AES tool, IntelliMetric™, and by human raters. Data collection included administering the Texas version of the WritePlacer *Plus* test and obtaining scores assigned by IntelliMetric™ and by human raters. The research sample of 107 participants was drawn from a Hispanic serving institution in South Texas. A One-Way Repeated-Measures ANOVA was conducted to examine the difference between the AES mean score and human raters' mean score. Results of the test indicated that the mean score assigned by the AES tool IntelliMetric™ was significantly higher than the faculty human raters' mean score on WriterPlacer *Plus* test. This finding did not corroborate previous studies that reported non-significant mean score differences between AES and human scoring.

J·T·L·A

Automated Essay Scoring Versus Human Scoring: A Comparative Study

Jinhao Wang
South Texas College
Michelle Stallone Brown
Texas A & M University, Kingsville

Introduction

The efficiency of automated essay scoring (AES) holds a strong appeal to institutions of higher education that are considering using standardized writing tests graded by AES for placement purposes or exit assessment purposes. However, it is not clear to what extent AES can replace human raters in judging the quality of essay writing. Research to date has mainly been conducted by testing agencies that market AES for commercial purposes. Companies such as Vantage Learning and ETS Technologies have published research results that demonstrate strong correlations and non-significant differences between AES and human scoring. However, the validity of AES tools is still a debatable issue. Some researchers criticized AES tools for their "over-reliance on surface features of responses, the insensitivity to the content of responses and to creativity, and the vulnerability to new types of cheating and test-taking strategies" (Yang, Buckendahl, and Juszkiewicz, 2002, p. 393).

Other researchers also suspected that the reported high percentage of agreement between AES and human raters might be due to the "interrelatedness of different elements in naturally occurring compositions; writers who produce well-organized passages also use a rich vocabulary and carefully revise mechanics" (Calfee, 2000, p. 35). In other words, AES was perceived as grading indirect features of writing that happened to correlate with the fine qualities of writing, so it is questionable whether AES could effectively rate essays that were "well-organized, but with poor mechanics or strong vocabulary but with lots of misspelling" (Calfee, 2000, p. 35).

Critics of AES also worried about the consequences of machine grading, which they believed would send the wrong message to students that writing was not important since the audience of student writers was

replaced by a machine. In addition, students would be focusing on writing the formulaic essay "that matches the computer's highest-score algorithm" (Baron, 2005, p. B14). The same concern was expressed by other scholars who argued that if students were asked to write to the machine rather than to human beings, they might think of writing as a "formal display" – just a "demonstration," not "words that might have an impact on another person and in some small way change the world" (Herrington and Moran, 2001, p. 496–497).

Many composition scholars also fear the negative impact of using AES tools on writing instruction. They believe that the AES approach conveys the message that "writing consists of discrete stylistic components that operate independently of communicative contexts," which means writing instructors can "revert to workbook exercises in vocabulary and complex sentences" instead of reading students' essays (Fitzgerald, 1994, p. 16). Scholars also criticized AES as endorsing counting rather than meaning making. They pointed out that AES "violates what effective teachers know about writing and assessment" (Cheville, 2004, p. 49). All in all, many composition scholars are convinced that AES tools cannot simulate the writing instructors' assessment process because human assessment of writing involved "relativities of reading," "multiple subjectivities," and "sophisticated intellectual operations" (Anson, 2003, p. 236). In other words, automated essay scoring is viewed as having a negative impact on writing instruction.

Evidently, while proponents of AES use validation studies to demonstrate the validity and effectiveness of AES, critics of AES still hold legitimate concerns. At the core of the debate is the issue of whether automatic scoring tools can indeed replace human raters in judging the qualities of writing valued by writing instructors. So far, very few studies have been conducted by independent researchers and users of AES. Institutions that are in the process of making decisions about whether or not to adopt AES for the benefit of its efficiency are left with little impartial research on which to base their decisions. Thus, it is imperative that more research be carried out to confirm and shed new light on the existing studies of AES validity and values. The current study is an attempt to fulfill this need.

The Purpose of the Study

The purpose of the current study was to investigate the validity and usefulness of automated essay scoring for large-scale placement tests by comparing the performance of AES with that of human raters in assigning group mean scores. Specifically, the researcher examined the performance of one automated essay scoring tool – IntelliMetric™, which was used to score an online standardized writing test – WritePlacer *Plus*. This test was administered to a group of Developmental Writing students in a two-year college in South Texas. Since the majority of the participants were Hispanic, this population represented a different population from the one whose essays served to train the scoring model of IntelliMetric™; therefore, examining how well IntelliMetric™ can be applied to scoring different population's writing might shed light on the generalizability as well as validity of IntelliMetric™. Furthermore, the study might produce research evidence that would contribute to the ongoing dialogue about the implication and usefulness of AES tools in writing assessment and writing instruction.

Review of Literature

An Overview of Automated Essay Scoring

As a relatively young field, AES has only a 40-year history. Ellis Page is generally regarded as the pioneer of AES (Bereiter, 2003; Kukich, 2000; Wresch, 1993). In 1966, Page designed a computer grading program named Project Essay Grader (PEG). Utilizing the statistical capabilities of computers, Page (1966) looked for the kind of textual features that could be extracted by computers from the texts and then applied multiple linear regression to "determine an optimal combination of weighted features that best predicted the teachers' grades" (Kukich, 2000, p. 22). Some of the features he identified as having predictive power included "word length, essay length in words, number of commas, number of prepositions, and number of uncommon words – the latter being negatively correlated with essay scores" (Kukich, 2000, p. 22). Page believed the computer extractable predictive features "approximated" the intrinsic features valued by human raters, so he termed these features as "proxes;" he then termed the intrinsic features valued by human raters as "trins" (Wresch, 1993, p. 46).

In 1968, Page published the results of a study he conducted for comparing his PEG rating of student essays with human raters. The multiple *R* correlation between PEG scores and teachers' scores was .78 whereas the multiple *R* correlation between two or more teachers was .85 (Kukich,

2000). This study utilized 30 proxes, among which seven correlated significantly to positive human ratings.

Although Page's pioneering work seemed promising, AES tools did not gain popularity for the next two decades. Page did, however, spark more research interests in the AES arena. As reported by Wresch (1993), in the 1970s, two researchers, Henry Slotnick and Patrick Finn, advanced AES research by experimenting with different approaches. Whereas Slotnick reversed Page's approach by identifying trins first and then organizing the proxes around trins, Finn looked for the correlation between the low frequency words and the quality of writing (Wresch, 1993).

The 1980s saw a change of direction from scoring essays to providing feedback on student essays. The Writer's Workbench tool (WWB) developed by AT&T was designed to provide feedback to writers in terms of "spelling, diction, and readability" (Kukich, 2000, p.23), and another revision tool similar to WWB called Writer's Helper (WH) was developed by Conduit to help writers check for word frequency, sentence variety, transition word, and paragraph development. In a 1990 study, Reed found that WH could help improve students' writing if students utilized the tool for revision. The experimental group that used WH earned an average essay score of 5.5 out of 6 whereas the control group earned an average score of 3.9 (Reed, 1990).

Both of these tools were more advanced than Page's original AES tool in that they looked for "markers of coherence," and inferred "style" (Wresch, 1993, p. 53). In addition, the idea of correlating word choice and readability levels to writing quality influenced AES researchers to come. In the early 1990s, two attempts were made to further advance AES research. One was Hal Hellwig's effort to design an AES tool to grade business writing. The other was the development of an AES tool for the Alaska Assessment Project (Wresch, 1993).

Hellwig (1990) used the idea of Semantic Differential Scale (SDS) – a scale formed by the "feel" of 1,000 commonly used words – to evaluate the quality of writing. Numerical ratings between -3 to +3 were assigned to each word based on three values: potency, evaluation, and activity (with a +3 representing the most powerful, most positive, and most active values, and a -3 representing the least powerful, least positive, and least active values). Hellwig's research opened up the possibility of correlating automated rater judgments "with subjective judgments founded on word choice" (Wresch, 1993, p. 52).

Influenced by Hellwig's approach, the Alaska Assessment Project administrators McCurry and McCurry (1992) developed a tabulation program that was based on the detection of textual features and variables that appeared to increase as students moved up to higher grade levels. The list of features and variables was an expansion on Page's list, incorporating "Fogg readability" as well as "Flesch readability" (both were readability indexes used to determine the reading levels of any text) in addition to other usages of words. This approach yielded a better result than Page's PEG, demonstrating higher correlations between the computer detection of textual features and human raters' holistic scores (Wresch, 1993, p. 54).

These research projects carried out in the early 1990s paved the way for a more advanced design of AES in the late 1990s. In addition, the advancement of natural-language processing (NLP) and information retrieval (IR) also enabled researchers to look for new approaches to extract measures that directly correlated to writing quality. During the late 1990s, three major AES devices were developed. They were Intelligent Essay Assessor by Pearson Knowledge Technologies, *e-rater* by Educational Testing Service (ETS), and IntelliMetric™ by Vantage Learning (Kukich, 2000).

The *e-rater* engine, originally named Computer Analysis of Essay Content, was first designed to grade the Analytical Writing Assessment part of the Graduate Management Admissions Test (GMAT). The grading criteria for evaluating this GMAT writing test included such qualities as "syntactic variety, topic content, and organization of ideas" (Kukich, 2000, p. 23). Researchers at ETS, headed by Jill Burstein, "hypothesized" groups of NLP and IR extractable linguistic features that might correlate with the GMAT grading criteria. For example, they "hypothesized" that syntactic variety could be measured by quantifying types of sentences and clauses used in the essays, and they could approximate values for these features by using "syntactic processing tools available in the NLP community" (Kukich, 2000, p. 23). They could also employ "vector space modeling techniques now common in IR" (Kukich, 2000, p. 23) to measure topic content. The *e-rater* instrument extracts more than 100 features and assigns values to each feature. *E-rater* then uses step-wise linear regression to decide on a scoring model that best predicts the human raters' scores (Kukich, 2000). After continuous improvement, *e-rater* is now able to grade not only GMAT with a high degree of agreement with human raters, but is also able to reliably score other types of essays (Kukich, 2000).

While ETS was focusing on developing *e-rater*, Landauer and Laham (2000) were designing another program called Latent Semantic Analysis (LSA). The underlying concept behind LSA is that "the aggregate of all the contexts in which words appear provides an enormous system of simultaneous equations that determines the similarity of meaning of words and passages to each other" (Landauer & Laham, 2000, p. 27). Every word and passage is represented as a "point" in "semantic space" and the similarity of meaning between two words and passages is determined by estimating their relative positions in the space (Landauer & Laham, 2000, p. 27). Using the LSA approach, Landauer and Laham (2000) developed Intelligent Essay Assessor (IEA), which not only scores essays in specific areas of study, but also serves as a learning tool. IEA provides feedback to students in three areas, namely, "content," "style," and "mechanics" (Landauer and Laham, 2000, p. 27). IEA has the advantages of being able to capture "transitivity relations and collocation effects among vocabulary terms, thereby letting it accurately judge the semantic relatedness of two documents regardless of their vocabulary overlap" (Kukich, 2000, p. 24–25).

As researchers at ETS and Pearson Knowledge Technologies were engaged in developing and applying AES tools such as *e-rater* and IEA, another company, Vantage Learning in affiliation with College Board, developed an AES tool called IntelliMetric™, which had undergone 10 years of an experimental stage and was released for commercial use in 1998 (Vantage Learning, 2001b). Vantage Learning researchers reported that they had blended artificial intelligence (AI) with natural language processing (NLP) and statistical technologies in developing IntelliMetric™ and that this AES tool is capable of analyzing more than 300 semantic, syntactic and discourse level features (Vantage Learning, 2001b). It functions by first reading a pool of essays with known scores determined by the expert raters. It then derives characteristics associated with the essays at different score levels. This process enables the establishment of a scoring model, which is then tested against another set of essays to confirm its effectiveness. Finally, the tested scoring model is applied to the scoring of new essays (Elliot, 2003). Because of the use of AI, IntelliMetric™ is believed to be able to identify "characteristics that human raters [are] likely to value and those they find poor" (Dikli, 2006, p.15).

Currently, AES researchers are focused on going beyond providing students with essay scores; they strive to "extract finer-grained features of writing," so as to give students and teachers useful feedback (Kukich, 2000, p. 25-26). Promising progress has been made in research studies that explore the correlation between writing quality and "lexical-grammatical errors," or "rough shifts," or "rhetorical relations" (Kukich, 2000, p. 26). However, more advances need to be made in AI and NLP research before these newly explored measures can be made operational. At the current

stage, the AES tools are still weaker than human raters in scoring the content of essays and in evaluating works written in non-testing situations (Warschauer & Ware, 2006).

Research on the Validity of AES Tools

Research on AES has mainly been conducted by the companies that developed the AES tools. Whereas most researchers have aimed at demonstrating how well AES correlated to human raters' scoring (see a summary in Warschauer & Ware, 2006), some researchers have also investigated the threats to the validity of AES, so as to improve the performance of AES tools (Burstein, Kukich, Wolff, Lu, and Chodorow, 1998; Powers, Burstein, Fowles, Chodorow, & Kukich, 2001). Furthermore, research efforts have been made to explore the effectiveness of AES for assessment in the classroom, thus expanding the potential of using AES to benefit writing instruction (Erickson, 2000; Riedel, Dexter, Scharber, & Doering, 2005).

As noted by Warschauer and Ware (2006), psychometric research on AES tools has generally supported the conclusion that the range of correlations between scores produced by AES tools and those assigned by human raters is comparable to the range of correlations between two human raters' scores. This conclusion is also corroborated by some of the research projects on PEG and *e-rater*. To examine the effectiveness of PEG for rating specific traits of writing, Page, Poggio, and Keith (1997) conducted a study, using a sample of 495 essays written by 12th graders for the writing assessment of the National Assessment of Educational Progress (NAEP) in 1988. Applying the Spearman-Brown prophecy test, they examined how well PEG would predict the average scores of eight raters as compared to the prediction rates of two, three, and four human raters. The results showed that PEG surpassed the prediction rates of two human raters on all the trait rating scores as well as on holistic scores, although when compared with the four-rater prediction rates, PEG prediction rates were lower on holistic scores and on two traits: style and mechanics.

Another study examining the validity of PEG in grading essay traits as well as holistic overall quality was conducted by Shermis, Koch, Page, Keith, and Harrington (2002). In this study, the validity of PEG was also tested by using Confirmatory Factor Analysis (CFA), which compared PEG scores to scores assigned by all possible pairs of six human raters. To avoid overlapping pairs, five different analyses were performed. The results showed that the standardized pattern coefficient for the human pairs ranged from .81 to .89, and the median coefficient was .86. However, for PEG, the coefficients ranged from .88 to .89 with a median coefficient of .89. These findings suggested that "the computer ratings of essays were at least as valid as pairs of human judges" (Shermis et al., 2002, p. 15).

Research on another AES tool, *e-rater*, also supported the validity of AES to a great extent. For example, Burstein, Kukich, Wolff, Lu, and Chodorow (1998) performed a study on the validity of *e-rater* when it was applied to scoring 500 Graduate Management Admissions Test (GMAT) essays and 200 Test of Written English (TWE) essays. The correlation analyses showed that *e-rater* had comparable correlation rates to those between the two human raters. Whereas the two human raters correlated with each other at rates ranging from .82 to .89 across the writing prompts, *e-rater* correlated with Rater 1 at rates ranging from .80 to .87 and with Rater 2 at rates ranging from .79 to .87 (Burstein, et al., 1998).

What distinguishes this study from other studies of AES validity is that it also made an attempt to look at the area of discrepancy – an area where the score difference went beyond one point difference. The rates of discrepancies between two human raters and between each human rater and *e-rater* were examined at each score level. The results showed that at the score level of 5 and 6, the rates of discrepancy between *e-rater* and each human rater were higher than the rates of discrepancy between the two human raters. Whereas the rate of discrepancy between two human raters was 8% at score level of 5 and 7% at score level of 6, the rate of discrepancy between *e-rater* and Rater 1 was 15% at score level of 5 and 34% at score level of 6. Similar discrepancy rates existed when comparing *e-rater* with Rater 2 (15% at score level of 5, and 31% at score level of 6) (Burstein, et al., 1998).

Research investigating the limits of AES tools was carried out by Powers, Burstein, Fowles, Chodorow, & Kukich in 2001. These researchers designed a study that specifically probed the threats to the validity of *e-rater*. For the purpose of this study, various writing experts and critics of AES were invited to produce writing responses to the Graduate Record Examination (GRE) writing prompts. These participants were encouraged to write in any way that they thought would "trick" the *e-rater* into overestimating or underestimating their essays. Furthermore, participants were asked to explain what discrepancies they would predict and what would cause those discrepancies. Once the essays were written, both human raters and *e-rater* scored these essays by using the holistic scoring guide designed for the GRE writing test (Powers et al., 2001).

Powers et al. (2001) found that 67% of the writing samples were correctly placed in the direction of score predictions (the mean scores assigned by *e-rater* were higher or lower than the mean scores given by the human raters, as predicted by the participants). Seventeen percent of the essays were placed in the wrong direction (their *e-rater* mean scores were higher or lower than human raters' mean scores when the predictions were the opposite). The other 17% of the essays had an *e-rater* rating

exactly the same as human rating although these essays were predicted to a higher or lower rating than *e-rater* rating (Powers et al., 2001). The researchers cited an example to demonstrate how *e-rater* could be tricked. One of the participants, a professor of computational linguistics, wrote a few paragraphs and copied them 37 times. The human raters gave his essay a score of 1, whereas *e-rater* gave him a score of 6 – the highest score (Powers et al., 2001).

In light of these findings, the researchers recommended that *e-rater* be used in conjunction with a human rater, and that further research focus on how to "identify excessively repetitive essays, as well as those that employ questionable logic" (Powers et al., 2001, p. 14). The post script of the study reported that ETS researchers have since then developed several filters to flag essays that had little lexical overlap or have excessively repetitive words. The flagged essays could then be sent to human raters for inspection (Powers et al., 2001).

Research on IntelliMetric™ has mainly focused on validating IntelliMetric™ as an effective AES tool. No studies have been done to show potential weaknesses of this AES tool. Since 1996, more than 120 studies have been carried out, and most of them utilized correlational study designs. Scores assigned by IntelliMetric™ were compared with those given by human raters to determine the agreement rates and correlational coefficient rates. In almost all of these studies, researchers reported high agreement rates and high correlational coefficient rates (Greer, 2002; Vantage Learning, 2001a; Vantage Learning, 2002). Most recently, Rudner, Garcia, and Welch (2006) also reported a correlational coefficient rate as high as .83 when examining the relationship between IntelliMetric™ scoring and human raters' scoring. Very few studies thus far have utilized comparative study designs. The few studies that did use comparative research designs reported non-significant differences between AES and human raters' scoring (Nivens-Bower, 2002; Vantage Learning, 2003).

Specifically, Nivens-Bower's (2002) comparative study was conducted at two New England community colleges. Thirty students from introductory writing classes at both colleges took the WritePlacer *Plus* test. Their essays were graded by IntelliMetric™ and then by two college faculty members from each college. The six-point scale WritePlacer rubric was used for scoring essays utilized by both studies. A paired-sample *t* test was run to compare the group means, and the Wilcoxon signed rank test was performed to examine the range of score frequencies. As reported by the researcher, the paired-sample *t* test revealed no significant differences in mean scores at the level of .05 and .01 (*t* value not reported). The Wilcoxon signed rank test showed no significant difference in the range of score frequencies (*W* value not reported). Based on these results, as well as the

high correlational coefficient rates, Nivens-Bower (2002) concluded that IntelliMetric™ "produced results consistent with what would be expected of faculty scores" (Nivens-Bower, 2002, p. 12).

In the comparative study conducted by Vantage Learning in 2003, IntelliMetric™ was applied to grading instructional literary analysis essays (Vantage Learning, 2003). The Vantage Learning researcher collected 400 written responses from 9th and 10th grade students in English classes (the school and its location were unspecified). These responses were split into two sets: 350 of them for training IntelliMetric™ and 50 for validation. All the responses were first graded by two human expert graders. Then IntelliMetric™ was trained by the 350 expert-scored essays, and finally the trained scoring model of IntelliMetric™ was put to use for scoring the remaining 50 essays. The results of significance testing showed no significant difference between the mean score assigned by the experts and the mean score assigned by IntelliMetric™ (t = .265, p < .05). The mean score averaged from human expert scoring was 2.98 with a standard deviation of 1.26 while the mean score averaged from IntelliMetric™ scoring was 3.18 with a standard deviation of 1.38. In addition, high agreement and high correlation coefficient rates were reported. Based on these results, the researcher concluded that IntelliMetric™ performance in scoring essays in instructional environments "exceeded the performance typically found with expert scorers" (Vantage Learning, 2003, p. 6).

One interesting study conducted by Murphy (2002) at Richland College utilized mean scores to examine the construct validity of WritePlacer *Plus* graded by IntelliMetric™ without focusing on the mean score differences. Instead, the mean scores were used to show how well they matched students' course levels. The sample of the study included 445 students enrolled in six English Skills course levels ranging from the lowest level to the highest level. Students took WritePlacer *Plus* in spring 2001. The average score of WritePlacer *Plus* for each level was compared to the students' level of course placement to determine whether the average WritePlacer scores sequenced in the same way as the course placement levels. Students placed at the higher level would receive better WritePlacer scores if IntelliMetric™ graded the test as expected. Both human raters and IntelliMetric™ graded on a scale of 1 to 8. The results supported the construct validity of the WritePlacer scores by showing that WritePlacer scores assigned by IntelliMetric™ followed an ordinal pattern moving from lowest to the highest, matching the levels of course placement (Murphy, 2002).

Overall, research on IntelliMetric™ has so far reported strong correlations and non-significant differences between IntelliMetric™ scoring and human raters' scoring. However, among the published studies, very few studies were focused on comparing group mean scores (Murphy, 2002; Nivens-Bower, 2002; Vantage Learning, 2003). It is evident that comparative research is greatly needed.

Research Question

The current study was guided by the following research question:

Is the group mean score assigned by IntelliMetric™ significantly different from the group mean score assigned by human raters on the standardized writing test WritePlacer *Plus*?

Methods and Procedures

Population and Sample

A sample of 107 developmental writing students was drawn from an accessible population of 498 developmental writing students from a Hispanic serving institution in South Texas. Of the 107 participants, 52% were male and 48% were female. The largest age group was the group ranging from 18 to 28 years old (83%), and the largest ethnic group was Hispanic or Mexican American group (98%). All participants were native English speakers taking the highest level of Developmental English course. However, due to their ethnic background, many of them might be bilingual, speaking both English and Spanish (percentage of participants being bilingual was not included in the data gathering).

Simple random sampling was used to select research participants. All available respondents' names were entered into the SPSS database alphabetically, and the Case Selection procedure was used to select a random sample of approximately 35% of the 284 cases, which yielded 107 cases.

Instrumentation

To collect data, WritePlacer *Plus* was administered to the participants of the current study. The instrument also included a grading rubric, which was used to guide human raters' scoring. WritePlacer *Plus*, a standardized test that measures entry-level college students' writing skills 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. For the purpose of the current study, the online WritePlacer *Plus* is utilized. When taking the test, examinees are expected to compose

a writing sample in response to a particular prompt, which elicits writing in the mode of persuasion (College Board, 2004). For the sake of confidentiality, the writing prompts for participants' essays used for this study were not revealed.

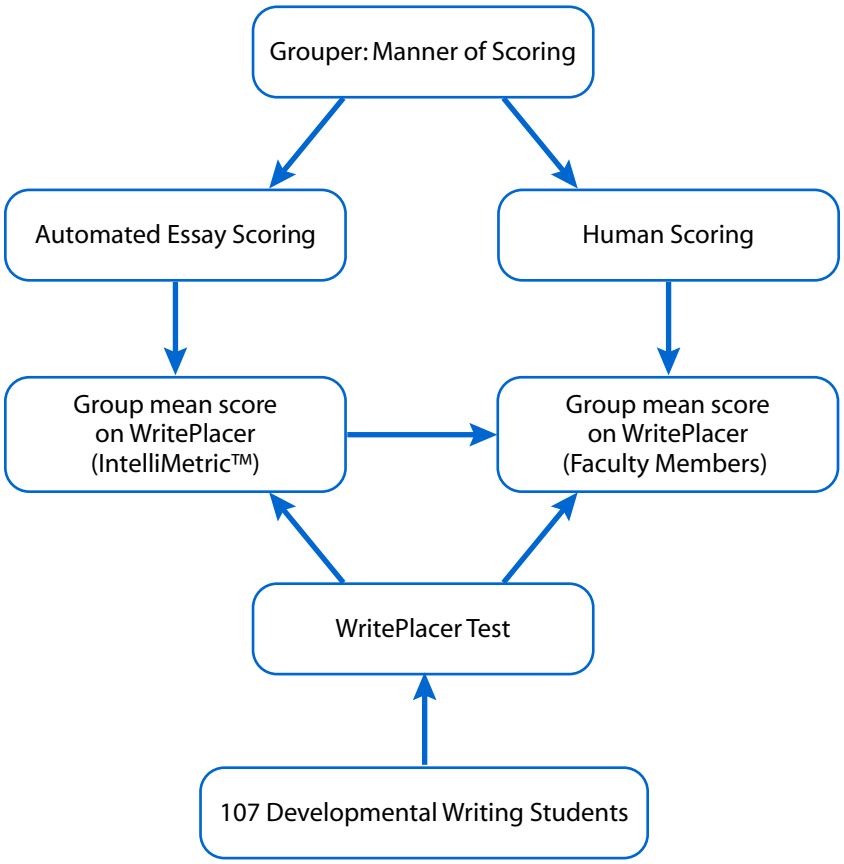
WritePlacer Score Point Description

Scores generated by IntelliMetric™ range from 2 to 8. Characteristics of writing samples at each score point are described in Appendix A (page 27). The examinees can interpret their WritePlacer scores by referring to this score point description.

Research Design

The current study utilizes the causal-comparative study design. Such a design involves "selecting two or more groups that differ on a particular variable of interest and comparing them on another variable or variables" (Fraenkel & Wallen, 2003, p. 371). In the current study, the researcher examined two groups, namely, the automated essay scoring (AES) group and the human raters' group on the WritePlacer *Plus* test (Figure 1).

Figure 1: Causal-comparative Design Model



The grouper or the preexisting factor that differentiated these groups was the manner of scoring – the automated essay scoring (AES) and the human scoring. The dependent variables on which the AES and human raters' scoring were compared were the group mean scores on the overall quality of the WritePlacer *Plus* writing samples holistically graded by IntelliMetric™ and by faculty human raters.

Data Collection

The Texas version of WritePlacer *Plus* was administered to the participants in the spring semester of 2006. The participants' WritePlacer *Plus* essays were scored instantly by IntelliMetric™ and then retrieved by the researcher. The retrieved essays were then scored by two trained faculty members. These two faculty members did not proctor the test, nor did they teach this group of participants.

Data Analysis

For the causal-comparative research design in the current study, the independent variable was the method of scoring with two levels, namely, the IntelliMetric scoring and the faculty human raters' scoring. The dependent variables were the group mean scores of the holistic scores assigned by the two grading methods. The one-way repeated-measures ANOVA was selected to test the significance of the group mean differences.

To conduct the one-way repeated-measures analyses, the researcher ran the test of main effect, using General Linear Model Repeated Measures procedures. The effect size and the descriptive statistics were also computed.

Results

Descriptive Statistics for the Overall Holistic Scores

First, the SPSS Explore procedure was run to examine the normality of the data. For the overall holistic scores assigned by the two scoring methods, namely, the IntelliMetric™ scoring of WritePlacer *Plus* (GRME1) and the human rater (faculty) scoring of WritePlacer *Plus* (GRME2), all 107 cases had valid scores, with no outliers. Table 1 displays the mean, median, and standard deviation for each scoring method.

Table 1: Descriptive Statistics for the Two Sets of Overall Holistic Scores, N = 107

Variables	M	Median	SD
GRME1 (WritePlacer AES)	5.98	6.00	.87
GRME2 (WritePlacer Human)	5.22	5.00	.96

Next, the Frequencies procedure was run to examine the distribution of the overall holistic scores in percentages, as assigned by the two scoring methods. For ease of comparison and discussion, the researcher presented the frequency tables by three score ranges, namely, by passing scores, borderline passing scores, and failing scores, as indicated in Table 2, Table 3 (next page), and Table 4 (next page). According to the Texas Higher Education Coordinating Board, for all Texas Success Initiative (TSI) approved writing assessment instruments, including WritePlacer *Plus*, a score of 6 and above indicates passing, a score of 5 indicates borderline passing (an essay score of 5 plus a score of 70% and above on the multiple-choice portion of the writing test indicates passing, whereas an essay score of 5 plus a score below 70% on the multiple-choice portion of the writing test indicates failing), and a score of 4 and below indicates failing (Texas Higher Education Coordinating Board, n.d.).

Table 2: Frequency Table for Scores of 6 and Above (Passing)

Variables	*n*	%
GRME1 (WritePlacer AES)	79	73.8
GRME2 (WritePlacer Human)	44	41.2

Table 3: Frequency Table for Scores of 5 (Borderline Passing)

Variables	*n*	%
GRME1 (WritePlacer AES)	25	23.4
GRME2 (WritePlacer Human)	34	31.8

Table 4: Frequency Table for Scores of 4 and Below (Failing)

Variables	*n*	%
GRME1 (WritePlacer AES)	3	2.8
GRME2 (WritePlacer Human)	29	27.1

As can be observed from the frequency tables, the AES tool, IntelliMetric™, assigned scores with a much higher passing rate (73.8%) than the pass rate assigned by faculty human raters (41.2%). In contrast, faculty human raters assigned scores with a higher borderline passing rate (31.8%) than the borderline passing rate assigned by IntelliMetric™ (23.4%). As far as the failing rate is concerned, IntelliMetric™ scores indicated a very low failing rate (2.8%), as compared with failing rate assigned by the faculty raters (27.1%).

Research Question

To address the research question, the following null hypothesis was tested:

> H_0 – There is no statistically significant difference between the group mean score assigned by IntelliMetric™ and the group mean score assigned by human raters on the standardized writing test WritePlacer *Plus*.

To evaluate this null hypothesis, a one-way repeated-measures ANOVA was conducted with the factor being the grading methods (IntelliMetric™ holistic scoring and faculty human raters' holistic scoring). The dependent variables were the numeric scores given by these two scoring methods. The means and standard deviations for these two sets of scores are presented in Table 1. The results for the ANOVA indicated a significant overall effect of grading methods, Wilks's $\Lambda = .712$, $F(1, 106) = 42.85$, $p < .01$, multivariate $\eta^2 = .288$ (Table 5, next page). Since the within-subjects factor had only two levels, no follow-up paired-samples t tests were needed.

The statistically significant effect of the grading methods indicated that the mean score assigned by IntelliMetric™ is significantly higher than human raters' mean score on the WritePlacer Plus test.

Table 5: Results of the One-Way Repeated-Measures ANOVA

Source	*df*	Wilks's Λ	*F*	η^2
Grading Methods (WritePlacer AES, WritePlacer human)	1	.712	42.85*	.288
Error *df*	106	—	—	—

*$p < .01$.

Interrater Reliability Analyses

To ensure interrater reliability, the two faculty raters received holistic scoring training from both the NES training center and from local trainers. In addition, calibrations took place prior to each grading section. For all holistic scores assigned, the two raters had no discrepancies that were two points or more apart. Intraclass correlation coefficients, using Two-Way Mixed-Effect model and Consistency definition, were computed to determine the level of interrater reliability between the two raters' overall holistic scores.

According to Garson (2006), the Intraclass Correlation Coefficient (ICC) statistic calculates the ratio of between-group variance to total variance. The Two-Way Mixed- Effect model in SPSS regards the raters as a fixed effect rather than random effect, whereas the students' essay scores in the sample are seen as a random effect. The ICC coefficient values for Two-Way Mixed-Effect are the same as the Two-Way Random- Effect model, except that in the Two-Way Mixed-Effect model, the ICC coefficients are not generalizable beyond the given raters. The selection of Consistency definition in SPSS means that ICC examines whether the two raters' scores are highly correlated; the similarity of the relative ratings rather than absolute agreement is the focus of the analysis.

For the current study, the result of ICC calculated by using the Two-Way Mixed-Effect model and Consistency definition in SPSS (Version 12.0.1) indicated an ICC value of .62, an acceptable level of interrater reliability for the overall holistic scoring, using .60 as the cut-off value for acceptability level (McGraw & Wong, 1996).

Discussion and Conclusion

The group mean score comparisons revealed significant differences between the overall holistic mean scores assigned by IntelliMetric™ (M = 5.92, SD = .87) and those scores given by faculty human raters on WritePlacer *Plus* (M = 5.22, SD = .97). IntelliMetric™ assigned significantly higher mean score than human raters' mean score. Few researchers, thus far, have examined differences in mean scores assigned by IntelliMetric™ and human raters. Out of the few published studies that did examine group mean scores (Murphy, 2002; Nivens-Bower, 2002; Powers, Burstein, Fowles, & Kukich, 2001; Vantage Learning, 2003), only two studies investigated the group mean differences, and the results showed no statistically significant difference in mean scores. While Nivens-Bower's study didn't report the t value, Vantage Learning's (2003) study reported a t value of .265 (p > .05). However, even in this latter case, though statistically non-significant, IntelliMetric™'s mean score (M = 3.18, SD = 1.38) was still higher than human raters' mean score (M = 2.98, SD = 1.26). It appears that IntelliMetric ™ tends to assign higher scores than do human raters. Furthermore, the descriptive statistics also indicated that IntelliMetric™ assigned a much higher passing rate (Table 2, page 17) and a much lower failing rate (Table 4, page 18).

The finding of a significant difference between IntelliMetric™ scoring method and human scoring method as demonstrated by the current study was also supported by the results of a larger research project (Wang, 2006), from which the current study was drawn. This larger research project included a correlational study design, examining correlations among three sets of overall holistic scores, which included those assigned by IntelliMetric™, by faculty human raters on WritePlacer *Plus*, and by NES human raters on a second writing test. The same group of participants who took WritePlacer *Plus* also took the second writing assessment. By adding the participants' performance on the second writing assessment to the study, the construct validity of IntelliMetric™ was tested.

The results of correlational analyses, using the nonparametric Spearman Rank Correlation Coefficient tests, indicated that no statistically significant correlations were present between the IntelliMetric™ overall holistic scores and faculty human raters' overall holistic scores (r_s = .11, p < .017), nor was there a significant correlation between the IntelliMetric™ overall holistic scores and the overall holistic scores of the second writing test (r_s = .04, p < .017). On the other hand, there was a statistically significant correlation between the overall holistic scores assigned by the faculty human raters and the overall holistic scores given by the NES human raters on the second writing test (r_s = .35, p < .017), a coefficient of medium size, according to Green's and Salkind's (2005) standard (Wang,

2006). It appears that the faculty human raters and NES human raters score more consistently with each other, whereas IntelliMetric™ scores less consistently with either faculty human raters or NES human raters.

In general, the results from the current study do not support findings published by Vantage Learning (Nivens-Bower, 2002; Vantage Learning, 2003). The discrepancies between the findings of the current study and those by the previous studies may be due to various factors, such as the difference in population and difference in the types of writing instruction. For instance, Nivens-Bower's (2002) study that showed non-significant mean score difference was conducted at two New England community colleges. Participants involved in this study were students taking introductory writing classes. In contrast, participants for the current study were from South Texas and were enrolled in the Developmental English classes. These participants were largely Hispanic, having different linguistic and cultural background from the New England students. Their writing would conceivably assume the idiosyncrasy of their unique background. In addition, writing instruction for an introductory college writing class might be different from a Developmental English writing instruction in terms of curricular and instructional goals, which might also impact how students write. Due to the linguistic, cultural, and instructional idiosyncrasies, the IntelliMetric™ scoring model may not be generalizable to student populations different from the population whose essays served as the training model for IntelliMetric™. The finding of significant mean score difference between IntelliMetric™ and human raters from the current study calls the generalizability of IntelliMetric™ into question, demonstrating the need for more studies investigating the generalizability of IntelliMetric™ across different student populations.

The tendency of IntelliMetric™ in assigning higher scores than human raters, if confirmed by future studies, may also mean that scores assigned by IntelliMetric™ may not be able to serve as an accurate placement instrument because students who are not college ready could be given a passing score by IntelliMetric™ and be placed at a level where they may experience difficulties in succeeding in their course work. As we know, a valid assessment tool should be able to identify whether students are "ready for a specific level of instruction" (Huot, 2002, p. 148). Assigning a higher score to a student's essay than what the essay qualifies directly impacts how well the student can perform in the course; therefore, it is important that more studies be conducted to assess if IntelliMetric™ does have the tendency to assign unwarranted high scores.

Based on these finding and implications, the researcher recommends that further studies be conducted to determine the validity and generalizability of AES tools. Topics should include experimental studies that investigate which surface features impact the AES tools' assigning high scores and comparative studies that examine the mean score differences across AES mean scores and human raters' mean scores, all on the same student writing samples. There is also a need to examine the effect to which differences in the characteristics of the training sample influence the scores awarded to a sample that has different characteristics. Finally, qualitative studies should be conducted to analyze essays that receive AES scores with a two-point discrepancy from human raters' scores.

In the interim, assessment programs should consider using multiple assessments, which include not only timed writing samples assessed by human raters, but also students' writing portfolios and advising/counseling processes. In addition, assessment programs should be flexible in allowing students to retest at any time. A process for validating and evaluating the assessment approach should also be in place to maximize the accuracy of placement and to generate knowledge about students' learning needs, curricular needs, and instructional needs.

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.

Baron, D. (2005, May 6). The College Board's new essay reverses decades of progress toward literacy. *The Chronicle of Higher Education*, P. B14.

Bereiter, (2003). Foreword. In M.D. Shermis & J.C. Burstein (Eds.), *Automatic essay scoring: A cross-disciplinary perspective* (pp. vii–x). Mahwah, NJ: Lawrence Erlaum Associates.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). *Computer analysis of essays*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA. Retrieved September 10, 2005, from http://www.ets.org/research/dload/ncmefinal.pdf

Calfee, R. (2000). To grade or not to grade. *IEEE Intelligent Systems 15*(5), 35–37.

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.

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment, 5*(1). Retrieved July 15, 2007, from http://www.jtla.org.

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 Erlaum Associates.

Erickson, J.D. (2000). Using keywords and computers to assess student writing (Doctoral Dissertation, Washington State University, 2000). *Dissertation Abstracts International, 61*, 3964.

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

Fraenkel, J.R., & Wallen, N.E. (2003). *How to design and evaluate research in Education* (5th ed.). New York: McGraw-Hill.

Garson, G.D. (2006). Reliability analysis. Retrieved September 9, 2006, from http://www2.chass.ncsu.edu/garson/pa765/reliab.htm

Green, S.B., & Salkind, N.J. (2005). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (4th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Greer, G. (2002). Are computer scores on essays the same as essay scores from human experts? In Vantage Learning, *Establishing WritePlacer Validity: A summary of studies* (pp. 10–11). (RB-781). Yardley, PA: Author.

Hellwig, H. (1990, March). *Computational text analysis for predicting holistic writing scores*. Paper presented at Conference on College Composition and communication, Chicago, IL.

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

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning*. Logan, UT: Utah State University Press.

Kukich, K. (2000). Beyond Automated Essay Scoring. *IEEE Intelligent Systems, 15*(5), 22–27.

Landauer, T.K., & Laham, D. (2000). The Intelligent Essay Assessor. *IEEE Intelligent Systems, 15*(5), 27–31.

McCurry, N., & McCurry, A. (1992). Writing assessment for the twenty-first century. *Computer Teacher, 19*, 35–37.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

Murphy, S. (2002). The relationship between WritePlacer *Plus* scores and course placements at Richland College. In Vantage Learning, *Establishing WritePlacer Validity: A summary of studies* (pp. 13–14). (RB-781). Yardley, PA: Author.

National Evaluation Systems. (2005). *THEA faculty manual: A guide to THEA test results*. Retrieved November 24, 2005, from http://www.thea.nesinc.com/PDFs/THEA_FacultyManual.pdf

Nivens-Bower, C. (2002). Faculty-WritePlacer *Plus* score comparisons. In Vantage Learning, *Establishing WritePlacer Validity: A summary of studies* (p. 12). (RB-781). Yardley, PA: Author. Norusis, M. J. (2004). SPSS 12.0 guide to data analysis. Upper Saddle River, NJ: Prentice Hall.

Norusis, M.J. (2004). *SPSS 12.0 guide to data analysis*. Upper Saddle River, NJ: Prentice Hall.

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

Page, E.B., Poggio, J.P., & Keith, T.Z. (1997, March). *Computer analysis of student essays: Finding trait differences in student profile*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED411316).

Powers, D.E., Burstein, J.C., Fowles, M.E., & Kukich, K. (2001, March). *Stumping e-rater: Challenging the validity of automated essay scoring*. (GRE Board Research Report No. 98-08bP). Princeton, NJ: Educational Testing Service. Retrieved October 15, 2005, from http://www.ets.org/Media/Research/pdf/RR-01-03-Powers.pdf

Reed, W. (1990, April). *The effect of composing process software on the revision and quality of persuasive essays*. Paper presented at the Eastern Educational Research Association, Clearwater, FL.

Riedel, E., Dexter, S.L., Scharber, C., & Doering, A. (2005, April). Experimental evidence on the effectiveness of automated essay scoring in teacher education cases. Paper presented for the 86th Annual Meeting of the American Educational Research Association, Montreal, CA.

Rudner, L.M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric™ essay scoring system. *Journal of Technology, Learning, and Assessment, 4*(4). Retrieved January 6, 2007, from http://escholarship.bc.edu/jtla/vol4/4/

Shermis, M.D., Koch, C.M., Page, E. B., Keith, T.Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62*(1), 5–18.

Texas Higher Education Coordinating Board. (n.d.). THEA, ASSET, COMPASS and ACCUPLACER. Retrieved September 12, 2006, from http://www.thecb.state.tx.us/facts/cd/Page8.htm

Vantage Learning (2001a). *Applying IntelliMetric™ to the scoring of entry-level college student essays*. (RB-539). Retrieved September 1, 2005, from http://www.vantagelearning.com/content_pages/research.html

Vantage Learning (2001b). *IntelliMetric™ : From here to validity*. (RB-504). Retrieved September 1, 2005, from http://www.vantagelearning.com/content_pages/research.html

Vantage Learning (2002). *A study of expert scoring, standard human scoring and IntelliMetric™ scoring accuracy for statewide eighth grade writing responses*. (RB-726). Retrieved September 1, 2005, from http://www.vantagelearning. com/content_pages/research.html

Vantage Learning. (2003). *A comparison of IntelliMetric™ and expert scoring for the evaluation of literature essays*. (RB-793). Retrieved September 1, 2005, from http://www.vantagelearning.com/content_pages/research.html

Wang, J. (2006). An analysis of automated essay scoring versus human scoring (Doctoral dissertation, Texas A&M University – Kingsville, 2006). *Dissertation Abstracts International, 68*(02).

Warschauer, M. & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. Language Teaching Research, 10, 2, 1–24. Retrieved June 05, 2006, from http://www.gse.uci.edu/faculty/markw/awe.pdf

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.

Appendix A

Table 6: WritePlacer Score Point Description for IntelliMetric™ Scores

Score	Score Point Description
2	The writer attempts to address the topic, but language and style are inappropriate for the given audience, purpose, and/or occasion. There is often no clear statement of a main idea or point of view and there is confusion found in the writer's efforts in presenting supporting detail. Any organization that is present fails to present an effective sequence of ideas. The sentence structure, when presented in paragraph form, is ineffective and few sentences are free of errors. Adding to the confusion is the writer's inability or lack of care in making word choices. There are many errors in mechanical conventions of grammar, spelling, and punctuation.
3	The writer is largely unsuccessful at communicating a main idea or point of view, and there is little evidence of an organizational structure. Ideas lack focus and development and there are many errors in mechanical conventions of usage, sentence structure, grammar, spelling, and punctuation.
4	A partially developed writing sample in which the characteristics of effective written communication are only partially formed. Statement of purpose is not totally clear and although a main idea or point of view may be announced, continued focus on the main idea is not evident. Development of ideas by the use of specific supporting detail and sequencing of ideas may be present, but is incomplete or unclear. Paragraphs are composed of sentences poorly structured with contain noticeable and distracting errors. The writer also exhibits poor precision in the use of grammatical conventions including poor word choice, poor usage, poor spelling and punctuation.
5	A writing sample that only partially communicates a message to the specified audience. The purpose may be evident but only partially formed. Focus on the main idea is only partially evident. The main idea is only partially developed with limited supporting details. While there is some evidence of control in the use of mechanical conventions such as sentence structure, usage, spelling and punctuation, some distracting errors may be present.
6	An adequately formed writing sample that attempts to communicate a message to a specified audience. Though the purpose of the writing sample may be clear, the writer's attempts to develop details may not be fully realized. The writer's organization of ideas may be characterized by a lack of specificity and/or incomplete development of ideas in effective sequence. Sentence structure within paragraphs is adequate though minor errors in sentence structure, usage, and word choice are evident. There are also errors found in the use of mechanical conventions such as spelling and punctuation.
7	A very good writing sample that substantially communicates a whole message to a specified audience. A purpose and focus is established, but may only be partially developed. An organizational pattern is evident, but is only partially fulfilled. 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.
8	A well-formed writing sample that effectively communicates a whole message to a specified audience. The writer maintains unity of a developed topic throughout the writing sample, and the writer establishes a focus by clearly stating a purpose. The writer exhibits control in the development of ideas and clearly specifies the supporting detail. The sentence structure is effective and free of errors. There is precision and care reflected in usage and choice of words as well as evidence of mastery of mechanical conventions such as spelling and punctuation.

Note. From ACCUPLACER coordinator's guide by College Board, 2004, p. 15.

Author Biographies

Jinhao Wang is the Chair of Developmental English Department at South Texas College. She obtained her doctorate in Educational Leadership from the Joint Doctoral Program in Educational Leadership at Texas A&M University – Kingsville and Corpus Christi. She has extensive experience in standardized testing, curriculum development, and community college teaching. Her research interests are primarily in the areas of standardized testing, English curriculum design and alignment, teacher effectiveness, and leadership in higher education. Dr. Jinhao Wang can be contacted at jinhao_wang@yahoo.com.

Michelle Stallone Brown is an Assistant Professor of Education in the Department of Educational Leadership and Counseling and Coordinator of the Doctoral Program in Educational Leadership at Texas A&M University – Kingsville. She obtained her doctorate in Educational Leadership from the Joint Doctoral Program in Educational Leadership at Texas A&M University – Kingsville and Corpus Christi. She has had the experience of serving as a statistical analyst for the City of Corpus Christi and performed statistical consulting work for the City of Corpus Christi. Her current research interests are primarily in the areas of No Child Left Behind policies, standardized testing, and student achievement. Dr. Michelle Stallone Brown can be contacted at michelle.stallone@tamuk.edu.

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College