To Aggregate or Not? Linguistic Features in Automatic Essay Scoring and Feedback Systems

by Scott A. Crossley, **Kristopher Kyle**, and **Danielle S. McNamara**

This study investigates the relative efficacy of using linguistic micro-features, the aggregation of such features, and a combination of micro-features and aggregated features in developing automatic essay scoring (AES) models. Although the use of aggregated features is widespread in AES systems (e.g., e-rater; Intellimetric), very little published data exists that demonstrates the superiority of using such a method over the use of linguistic micro-features or combination of both micro-features and aggregated features. The results of this study indicate that AES models comprised of micro-features and a combination of micro-features and aggregated features outperform AES models comprised of aggregated features alone. The results also indicate that that AES models based on micro-features and a combination of micro-features and aggregated features provide a greater variety of features with which to provide formative feedback to writers. These results have implications for the development of AES systems and for providing automatic feedback to writers within these systems.

Introduction

Developing automatic essay scoring (AES) systems to assess writing quality is an important component of standardized testing (Attali & Burstein, 2006), intelligent tutoring systems that focus on teaching writing (e.g., Writing-Pal, McNamara et al., 2012), and, in some respects, classroom teaching (Warschauer & Grimes, 2008), although many scholars and teachers voice concern over the use of AES systems in the classroom (Condon, 2013; Deane, 2013; Elbow & Belanoff, 1986; Haswell, 2006; Haswell & Ericsson, 2006; Herrington & Moran, 2001; Huot, 1996; Perelman, 2012, 2014; Wardle & Roozen, 2013). There are generally two approaches to automatically scoring writing quality: a micro-feature approach and an aggregated feature approach. A micro-feature approach focuses on using individual variables related to text length, lexical sophistication, syntactic complexity, rhetorical elements, and essay structure to predict human scores of writing quality. These variables have proven to be strong indicators of essay quality (McNamara, Crossley, & Roscoe, 2013) and can lead to relatively straightforward interpretations. However, potential concerns with a micro-feature approach are that statistical models of essay quality may contain redundant variables (e.g., a model may contain multiple variables of lexical sophistication) and that the developed models may not be stable or representative of larger linguistic constructs. Thus, some scoring engines such as e-rater (Burstein, 2003; Burstein, Chodorow, & Leacock, 2004) and Intellimetric (Rudner, Garcia, & Welch, 2005, 2006; Shultz, 2013) use aggregated features to assess writing quality. These aggregated features are generally average combinations of theoretically related micro-features (e.g., an aggregated score for mechanics may be based on the combination of the micro-features spelling, capitalization, punctuation, fused words, compound words, and duplicated words). However, it is yet to be seen whether micro-features, aggregated features, or a combination of both are most predictive of human judgments of essay quality. Thus, our primary goal in this study is to investigate the potential for scoring models based on micro-features, aggregated scores, and a combination of both to assess human ratings of essay quality. We also explore the practicality of using such models in delivering feedback to users in automatic writing evaluation (AWE) systems.[1]

To investigate the potential for using micro-features, aggregated scores, and a combination of each in AES systems, we developed statistical scoring models using linguistic indices reported by Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014), Linguistic Inquiry and Word Count (LIWC, Pennebaker, Booth, & Francis, 2007) and the Writing Assessment Tool (WAT, Crossley, Roscoe, & McNamara, 2013; McNamara et al., 2013). To develop aggregated scores, we took a novel approach and used a principal component analysis (PCA) to create weighted component scores (though see Attali & Powers, 2008, and Somasundaran, Burstein, & Chodorow, 2014 for related approaches).[2] PCA is ideal for this purpose because it uses co-occurrence patterns to transform correlated variables within a set of observations (in this case a corpus of persuasive essays) to linearly uncorrelated variables called principal components. Correlations between the individual indices and the components themselves can be used as weights from which to develop overall component scores (i.e., an aggregated scoring model). Thus, unlike previous research, the aggregated scores used in this study are not based on theoretical selections of averaged microfeatures, but rather on weighted statistical co-occurrence. We use the derived component scores, the individual micro-features that inform these component scores, and a combination of both to assess associations between human ratings of essay quality and linguistic features and develop scoring models that can be used to inform AES systems. The variables that inform these models are discussed in terms of predictive power, salience, and potential for use in writing feedback.

Automated Essay Scoring and Writing Evaluation

Automated essay scoring (AES) is the use of computers to predict human ratings of essay quality. AES can be helpful in both classroom settings and in high-stakes testing by increasing reliability and decreasing the time and costs normally associated with essay evaluation (Bereiter, 2003; Burstein, 2003; Higgins, Xi, Zechner, & Williamson, 2011; Myers, 2003; Page, 2003). In the classroom, AES systems can provide students opportunities for evaluated writing practice in addition to the evaluation opportunities a teacher provides (Dikli, 2006; Page, 2003). In high-stakes testing, AES is being used for assessments such as the Graduate Record Exam (GRE) and the Test of English as a Foreign Language internet-Based Test (TOEFL iBT) in order to provide cost-effective and reliable scores on performance assessments (Dikli, 2006).

Despite advantages of AES, some scholars have voiced a number of criticisms concerning its use. These criticisms revolve around the idea that AES systems are not capable of scoring essays in the same manner as do human raters. AES systems, for example, are not able to attend to essential aspects of writing such as rhetorical effectiveness, argumentation, purpose, or audience (Condon, 2013; Deane, 2013; Haswell, 2006; Haswell & Ericsson, 2006; Herrington & Moran, 2001; Huot, 1996; Perelman, 2012). Additionally, the effectiveness of AES models has generally been limited to shorter essay types, such as those found on standardized tests, and have been less effective in scoring more authentic writing assessments (e.g., portfolios, Condon, 2013; Elbow & Belanoff, 1986; Wardle & Roozen, 2013).

A number of prominent AES systems are currently in use, including e-rater (Burstein, 2003; Burstein et al., 2004; Burstein et al., 2013), Intellimetric (Rudner et al., 2005, 2006; Schultz, 2013), Intelligent Essay Assessor (IEA, Landauer, Laham, & Foltz, 2003; Foltz, Streeter, Lochbaum, & Landauer, 2013), LightSIDE (Mayfield & Ros é, 2013) and Writing Pal (W-Pal, Crossley et al., 2013). Although variation exists in the details of how the scoring models produced by AES systems function, they are developed using the same general methods. The first step in creating a scoring model is to collect a large number of writing samples that have been given scores by human raters. Textual features that are of theoretical interest to the construct(s) being assessed and can be reliably identified by a computer program (e.g., grammatical accuracy, lexical sophistication, syntactic complexity, and rhetorical features) are identified. The textual features are then calculated for each essay, and statistical models are created that use textual features to predict human scores. These models can be specific to a writing prompt or generalized across a number of prompts (Attali & Burstein, 2006).

Automatic writing evaluation systems (AWE) are generally built upon AES technology, but go beyond simply providing a predicted essay quality score by giving feedback on student writing, ideally in a clear, formative manner. Despite the promise of AWE systems to provide focused, formative feedback, some students are less likely to trust a computer system than a teacher, limiting AWE effectiveness (Grimes & Warschauer, 2010). Additionally, many AWE systems are still focused more on summative than formative feedback (Roscoe, Kugler, Crossley, Weston, & McNamara, 2012). Furthermore, AWE systems may not always provide effective individualized support due to issues some writers have that are generally infrequent within a larger population of writers but frequent to a specific writer. Nonetheless, there is some evidence that AWE systems are perceived to be effective when used at the early stages of essay composition (Chen & Cheng, 2008).

Reliability and Accuracy of Automated Essay Scoring Systems

High levels of agreement have been reported in a number of studies between AES systems and human raters (Attali & Burstein, 2006; Landauer et al., 2003; Landauer, Laham, Rehder, & Schreiner, 1997; McNamara et al., 2013; Vantage Learning, 2003; Warschauer & Ware, 2006; Shermis & Hamner, 2013). This agreement has been reported with regard to both correlations and quadratic weighted Kappa, the latter of which is the current standard (e.g., Shermis & Hamner, 2013). The level of agreement between AES systems and human raters is similar to the level of agreement between two human raters. Correlations for between human raters and human raters and AES systems tend to range from r = .70 to r = .85 (Warschauer & Ware, 2006). Agreement with regard to quadratic weighted Kappa range from 0.60 to 0.85 (Shermis & Hamner, 2013). Some studies have reported higher levels of agreement between AES systems and human raters than between two human raters. Weigle (2010), for example, reported that the two human raters in her study achieved correlations ranging from .64 - .67, while the correlations between the averaged human scores and e-rater ranged from .76 - .81. Other systems have also reported high AES and human rater agreement such as Intellimetric (r= .83, Rudner et al., 2006) and W-Pal (r = .81, McNamara et al., 2013).

In addition to computing correlations to measure inter-rater reliability, true agreement between human raters or between human raters and an AES system is often reported using exact and exact/adjacent agreement. Exact agreement is the ratio of essays that are given the same score by two raters (human or AES) to those that are given different scores (usually expressed as a percentage). Exact/adjacent agreement is the ratio of essays that are given either the same scores or scores that only differ by one point. Exact and exact/adjacent agreement between human raters and AES systems has been reported to be comparable to the exact and exact/adjacent agreement between two human raters. For example, the e-rater system has been reported to achieve between 57-59% exact agreement with human raters, as compared to 56-60% exact agreement between two human raters (Attali, 2008; Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012). The same studies reported that exact/adjacent agreement between e-rater and human scores ranged from 98-99%, which was higher in each study than the exact/adjacent agreement between two human raters of 97-98%. Agreement between IntelliMetric scores and human rater scores has also been high (ranging from 42%-65% exact and 92%-100% exact/adjacent agreement, Rudner et al., 2006) as have those for W-Pal (ranging from 44-55% exact and 94-96% adjacent, Crossley, Kyle, Allen, Guo, & McNamara, in press; McNamara, et al., 2013).

Micro-features and Essay Quality

Most published investigations into the relationship between essay quality and textual micro-features have been in a second language (L2) context, likely because L2 writers can vary greatly in terms of linguistic production and L2 writing is dependent on both writing ability and language ability (Weigle, 2013). Many early studies did not focus specifically on AES systems, but rather on relationships between micro-features and writing quality. These early studies were important, however, in guiding the later development of AES systems, which occurred in the late 1990s. Micro-feature studies have reported that high-quality L2 essays tend to be longer (e.g., Ferris, 1994; Frase, Faletti, Ginther, & Grant, 1999; Reid, 1990; Guo, Crossley, & McNamara, 2013), use

longer words (Frase et al., 1999; Grant & Ginther, 2000; Reid, 1986, 1990; Reppen, 1994) that are less frequent, less familiar, and have fewer associations (Crossley & McNamara, 2012; Guo et al., 2013), and use more varied lexis (Crossley & McNamara, 2012; Engber, 1995, Grant & Ginther, 2000; Jarvis, 2002; Reppen, 1994). From the perspective of syntactic complexity, high-quality L2 essays tend to use the passive voice more often (Ferris, 1994; Grant & Ginther, 2000) than lower quality L2 essays. With regard to cohesion, findings have been mixed. In some studies, the presence of cohesive devices has been associated with higher quality L2 essays (e.g., connectives, Jin, 2001), while in more recent studies, the presence of cohesive devices has been associated with lower quality L2 essays (e.g., lexical overlap, semantic overlap, givenness, and connectives, Crossley & McNamara, 2012; Guo et al., 2013).

First language (L1) studies that have used micro-features to assess essay quality have found that high quality L1 essays and/or essays written by students at a higher grade level tend to include more words and word types (Crossley, Allen, Kyle, & McNamara, 2014; McNamara, et al., 2013; Witte & Faigley, 1981), longer words, (Crossley et al., 2014), less frequent words (Crossley et al., 2014; Crossley, Weston, McLain Sullivain, & McNamara, 2011; McNamara, Crossley, & McCarthy, 2010; McNamara et al., 2013), less frequent bigrams (Burstein et al., 2004; Crossley, Cai, & McNamara, 2012) and include greater lexical diversity (Crossley, Weston et al., 2011; McNamara et al., 2010, McNamara et al., 2013). With regard to word information, essays that are of higher quality or written by more advanced writers tend to be more concrete (Crossley, Weston et al., 2011), and more imageable, but contain fewer associations (McNamara et al., 2013). Higher quality/more advanced essays also tend to be more syntactically complex. For example, they tend to have more T-units per sentence (i.e., a main clause and dependent clause combined, Yang, Lu, & Weigle, 2015), T-units with more words (Witte & Faigley, 1981), and more clauses per T-unit (Yang et al., 2015), but fewer clauses (Beers & Nagy, 2009) that are longer (Yang et al., 2015). In addition, higher quality/more advanced essays tend to have more modifiers per word phrase, more modifiers per noun phrase (Crossley, Weston, et al., 2011), include more words before the main verb (McNamara et al., 2010), more prepositions, more instances of subordination, more passives (Connor, 1990), fewer base forms of verbs (Crossley, Roscoe, McNamara, & Graesser, 2011), and more negations (Crossley et al., 2014). In terms of cohesion, higher quality/more advanced essays also tend to include a higher ratio of given to new information (McNamara et al., 2013; Crossley, Roscoe, et al., 2011), less content word overlap (Crossley, Weston, et al., 2011), and fewer positive logical connectives (Crossley, Weston, et al., 2011), though some studies (e.g., Witte & Faigley, 1981) have found that high quality essays contain more cohesive ties. Advanced academic student writing also includes more approximate hedges, glosses, concessions, and expressions indicating contrasts than less advanced student writing (Aull & Lancaster, 2014). In general, the findings from these studies indicate that micro-features related to lexical sophistication, syntactic complexity, and, to some degree, cohesion can be used to distinguish high proficiency from low proficiency essays.

Aggregated Features and Essay Quality

In contrast to research concerning the relationship between micro-features and essay quality, less research explicitly discussing the relationship between aggregated features and essay quality has been published even though at least two prominent AES systems rely on aggregated features to inform scoring models: e-rater and Intellimetric. e-rater, for instance, uses aggregated feature scores to assign essay scores for the GRE and TOEFL exams. Depending on whether a prompt-specific or general scoring model is used, e-rater creates scores based on 7-8 aggregated scores or macro-features including *grammar*, *usage*, *mechanics*, *style*, *organization*, *development*, *lexical complexity*, and *topic-specific vocabulary usage* (Enright & Quinlan, 2010; Attali & Burstein, 2006; Burstein, Tetreault, & Madnani, 2013). Each macro-feature is comprised of a number of micro-features. The macro-feature usage, for example, is comprised of the micro-features *article errors*, *confused words*, *incorrect word forms*, *faulty comparisons*, and *non-standard verbs or word forms* (Enright & Quinlan, 2010). A review of the pertinent literature indicates that the micro-features that inform the aggregated scores discussed above are theoretically selected and that micro-feature counts are evenly averaged to calculate the aggregated score (Enright & Quinlan, 2010; Attali & Burstein, 2006; Burstein, Tetreault, & Madnani, 2013). Intellimetric (Rudner et al., 2006) also uses aggregated scoring models to assess essay quality. Over 300 micro-features are calculated for each essay, which are aggregated into five macro-features, including *focus and unity*, *organization*, *development and elaboration*, *sentence structure*, and *mechanics and conventions* (Dikli, 2006; Schultz, 2013). It is unclear how the microfeatures in Intellimetric are combined into aggregated scores. Despite the use of aggregated feature in AES systems such as e-rater and Intellimetric, and a large body of research describing the effectiveness of these tools (e.g., Shermis & Hamner, 2013), relatively little specific information regarding the micro-features used to create the aggregated models or the manner in which the features are aggregated has been published.

Method

This study investigated the potential for automated linguistic indices to be combined into component scores through a principal component analysis (cf., Attali & Powers, 2008; Somasundaran et al., 2014). Our goal was to develop aggregated constructs that might better assess writing quality (e.g., Attali & Burstein, 2006). The study also investigated whether these component scores alone are an improvement to only using individual linguistic indices (i.e., micro-features) or if a combination of micro-features and component scores can best explain human ratings of essay quality. To accomplish this, we conducted a principal component analysis (PCA) using a number of natural language processing indices reported for a corpus of essays to examine if these indices co-occur with one another. Based on co-occurrence, we developed component scores and tested these scores in a regression model to predict human judgments of essay quality. We then developed models of essay scoring using micro-features alone and

micro-features combined with the component scores to compare differences between the approaches. The development of component scores based on linguistic features may prove beneficial in providing summative feedback to users in an automatic writing evaluation (AWE) system. More importantly, because component scores report on larger linguistic constructs, they have the potential to provide more accurate formative feedback to students in AWE systems, which may assist them in successfully revising essays.

Corpus

We used a generic corpus comprised of 997 persuasive essays. All essays were written within a 25-minute time constraint. The essays were written by students at four different grade levels (9th grade, 10th grade, 12th grade, and college freshman) and on nine different persuasive prompts (Appendix A presents example prompts and assignments and Appendix B presents example essays). The students came from five different geographical regions in the United States. The majority of the essays were typed using a word processing system. A small number were hand written and were subsequently transcribed. Details for the corpus are provided in Table 1. We divided the corpus into a training set and a test set following a 67/33 split (Witten, Frank, & Hall, 2011). Thus, we had a training set of 673 essays and a test set of 304 essays. The training set was used to develop scoring models while the test set was used to assess how well the models worked on an independent data set.

Table 1: *Descriptive statistics for corpus*

Short prompt	n	Grade level	Region
Competition	64	9th-11th	Arizona
Competition	126	10th	Washington, D.C.
Fame	133	10th	Washington, D.C.
Fitting In	35	First-year college	Tennessee
Heroes	158	First-year college	Mississippi
Images	64	9th-11th	Arizona
Images	126	10th	Washington, D.C.
Memories	45	9th-11th	New York
Optimism	56	9th-11th	New York
Uniqueness	155	First-year college	Mississippi
Winning	35	First-year college	Tennessee

Human scores

Two expert raters with at least 2 years of experience teaching freshman composition courses at a large university rated the quality of the essays using a standardized SAT rubric (http://sat.collegeboard.org/scores/sat-essay-scoring-guide). The SAT rubric generated a rating with a minimum score of 1 and a maximum of 6. A high-scored essay, according to the rubric generated a strong point of view and demonstrated critical thinking. Linguistically, a high-scored essay was coherent with a clear focus, smooth progression of ideas, and strong organization. In addition, a highly scored essay used a variety of accurate and apt vocabulary, provided appropriate examples, reasons, and evidence, and demonstrated syntactic variety. Raters were informed that the distance between each score was equal. The raters were first trained to use the rubric with 20 similar essays taken from another corpus. Once they reached agreement of $r = .70$, they scored the essays in the corpus independently. The final interrater reliability for all essays in the current corpus was $r > .74$. Final weighted Kappa for the two raters was .739.[3] The raters displayed exact agreement in 59% of the cases. The raters displayed adjacent agreement 39% of the time. In total, exact-plus-adjacent ratings occurred 98% of the time between the expert raters. The mean score between the raters was used as the final value for the quality of each essay. The mean score for the essays was 3.03 and the median score was 3. The scores were normally distributed.

Natural Language Processing Tools

We selected 211 linguistic indices from three natural language processing (NLP) tools: Coh-Metrix (Graesser, McNamara, Louwerse & Cai. 2004; McNamara et al., 2014), Linguistic Inquiry and Word Count (LIWC, Pennebacker et al., 2007), and the Writing Assessment Tool (WAT, Crossley et al., 2013; McNamara et al., 2013). These indices measured text cohesion (e.g., incidence of

connectives, word overlap between sentences), lexical sophistication (e.g., word frequency, word familiarity, word concreteness), situational cohesion (e.g. causal, temporal, and spatial cohesion), syntactic complexity (e.g., phrasal and clausal complexity), n-grams accuracy and frequency (e.g., bi- and tri-grams), text relevance (e.g., overlap with prompt terms and key words in essays), rhetorical strategies (e.g., discourse cue words and terms), grammar (e.g., tense and aspect), psychological terms, (e.g., negative and positive emotional words, cognitive words), and personal terms (e.g., terms related to religion, money, and death). These tools and the indices they report on are discussed briefly below. We refer the reader to Crossley et al., (2013); McNamara et al. (2013), McNamara et al. (2014), and Pennebaker et al. (2007) for further information about the tools.

Coh-Metrix. Coh-Metrix is a computational tool that measures a number of text features related to syntactic difficulty, lexical sophistication, text structure, and cohesion. Coh-Metrix integrates lexicons, pattern classifiers such as topic sentence identifiers, part-of-speech taggers, syntactic parsers, shallow semantic interpreters such as Latent Semantic Analysis, and other NLP components that have been developed to analyze text characteristics. The indices reported by Coh-Metrix are primarily related to text difficulty and include indices of local and global cohesion. For instance Coh-Metrix uses Latent Semantic Analysis (Landauer, McNamara, Dennis, & Kintsch, 2007; McNamara et al., 2012) to measure semantic similarity between sentences and paragraphs, and to provide an index of text givenness (i.e., the amount of information that is recoverable or repeated from the preceding discourse, Hempelmann, Dufty, McCarthy, Graesser, Cai, & McNamara, 2005). Coh-Metrix also reports on lexical overlap indices between sentences (nouns, content words [nouns, verb, adjectives, and adverbs], word stems, and arguments [word, word stems, and pronouns]) along with a number of type-token ratio indices (D; Malvern, Richards, Chipere, & Durán, 2004; MTLD; McCarthy & Jarvis, 2010) and connective counts. The tool also reports on lexical sophistication (e.g., CELEX word frequency, average syllable per words, MRC indices related to imageability, concreteness and meaningfulness, Coltheart, 1981; WordNet indices related to hypernmy and polysemy, Fellbaum, 1998), situational cohesion (e.g., temporal, causal, and spatial cohesion), part of speech (POS) tags (nouns, verbs, adjectives, infinitives, prepositions, determiners) and phrasal tags (e.g., noun, verb, adjective, preposition, and adverb phrases), syntactic complexity (e.g., incidence of subordinate clauses, phrase length, mean number of words before the main verb, and a replication of the Biber tagger, Biber, 1988), and syntactic similarity (e.g., the uniformity and consistency of syntactic constructions in the text at the clause, phrase, and word level), and basic text measures (e.g., text, sentence, and paragraph number and length, type counts).

LIWC. LIWC reports on words and terms for grammatical variables, psychological variables (e.g., social, affective, and cognitive word incidence scores), and personal variables (leisure, work, religion, home, and achievement word incidence scores). Grammatical categories include pronouns, negation, function words, auxiliary verbs, articles, and words related to past, present, and future. Affective variables reported by LIWC relate to emotion words such as sadness, anxiety, and anger. Cognitive categories are related to certainty, inhibition, and inclusion/exclusion, while personal variable measure the incidence of words related to social items such as friends and family.

Writing Assessment Tool. WAT represents the state-of-the-art in computational tools and computes linguistic features specifically developed to assess student writing. These features include indices related to global cohesion (e.g., LSA semantic overlap between paragraph types such as essay introductions, body paragraphs, and conclusion and verb cohesion), lexical sophistication (e.g., academic words, vague nouns), topic development (e.g., key word use in essays), topic relevance (i.e., LSA semantic similarity between prompt and essay), n-gram use (e.g., bi-gram and tri-gram accuracy, frequency, and proportion scores, Crossley et al., 2012), and rhetorical features (e.g., amplifiers and emphatics [extremely, definitely], discourse connectors [conjuncts], and downtoners and hedges [slightly, somewhat, almost]).

Statistical Analysis

To compute the component scores, we adopted an approach similar to Graesser, McNamara, & Kulikowich (2011). We conducted a principle component analysis (PCA) to reduce the number of indices selected from the NLP tool to a smaller set of components, each of which is comprised of a set of related features. The PCA clustered the indices into groups that co-occurred frequently within the texts allowing for a large number of variables to be reduced into a smaller set of derived variables (i.e., the components).

For inclusion into a component, we set a conservative cut off for the eigenvalues to ensure that only strongly related indices would be included in the analysis. For inclusion in the analysis, we first checked that all variables were normally distributed. We then controlled for multicollinearity between variables (defined as $r > .90$) so selected variables were not measuring the same construct. After conducting the factor analysis, we then used the eigenvalues in the selected components to create weighted component scores. These component scores were then correlated against the human essay scores to examine potential associations with essay quality.

The component scores that demonstrated a significant correlation and at least a small effect size ($r > .100$) were then used in a regression analysis to model the human scores of essay quality. Those component scores that demonstrated significant correlations with the human scores in the training set were retained as predictors for the regression analysis. Using stepwise method, the component scores were next regressed against the holistic scores for the training set essays with the human scores as the dependent variable. The regression model weights were then applied to the essays in the test sets to predict the holistic scores. We

report on the *r* and R^2 values, which provide us with an estimate for the amount of variance in the human scores that the model explains and a measure to compare the results of the regression model to the human raters (i.e., comparing the *r* values between the human raters and the model). The model from this regression analysis was then applied to the held back essays in the test set to assess how well the model worked on an independent set of essays (i.e., how generalizable the model is to essays it was not trained on).

We also report exact and adjacent matches between the model and the human score. The premise behind such an analysis is that a score that is only off by one point (i.e., adjacent accuracy) is more acceptable than a score that is off by 2 or more points and, in human rating, generally scores that are one-point different are averaged, while scores that are two or more points off require adjudication (Attali & Burstein, 2006; Dikli, 2006; Rudner et al., 2006; Shermis, Burstein, Higgins, & Zechner, 2010). We rounded all half scores to convert them to integers. This rounding meant that all half scores were rounded up such that a 4.5 became a 5. For the exact scores, we calculated a quadratic weighted kappa, which allowed for comparisons between the human scores and the regression model scores. Similar regression methods were used for the micro-feature only analysis and then for the combined micro-feature and component score analysis.

Results

Assumptions for Principle Component Analysis

Of the 211 selected variables, 68 were not normally distributed and were removed from the analysis. The majority of these variables were bag-of-word counts taken from LIWC or the Biber tagger found in Coh-Metrix. In both cases, the words that informed the variables were highly infrequent in the essays leading to positively skewed distributions. Of the remaining variables, an additional 20 variables were removed because of strong multicollinearity with another variable. After controlling for both normal distribution and multicollinearity, we were left with 123 variables to include within the PCA.

Principle Component Analysis (PCA)

The PCA reported 32 components with initial eigenvalues over 1. Within the rotated components, there was a break in the cumulative variance explained between the ninth and tenth component. These nine components explained approximately 37% of the shared variance in the data for the rotated components. Considering this break, we decided on a 9-component solution when examining the PCA. Each of these components and the indices that inform them along with their weighted scores and correlations with human ratings are discussed below.

Component 1: Lexical and Nominal Simplicity. The first component (see Table 2) seemed to capture lexical and nominal simplicity. From a lexical simplicity standpoint, the component included more frequent words, shorter words, more frequent trigrams, more familiar words, more specific words, fewer academic words, and more social words (i.e., common words like *family*, *people*, and *talk*). From a nominal perspective, the component had lower lexical density, fewer nouns, fewer nominalizations, and more s-bars (i.e., more sentences with ellipsed nouns).

We calculated a correlation between the weighted scores for this component for each essay and the human ratings for essay quality. The correlation reported $r(997) = -.359$, $p < .001$, indicative of a moderate negative relation with essay quality. The correlation reflects the notion that essays with more frequent words and fewer nouns are scored lower by expert raters.

Table 2:*Component 1 indices and loadings: Lexical and nominal simplicity*

Index	Eigen loading
CELEX frequency (all words)	0.818
Average syllables per word	-0.809
CELEX frequency (content words)	0.748
Proportion spoken trigrams	0.735
Lexical density	-0.709
Word familiarity	0.648
Incidence of all nouns	-0.643
Hypernymy	-0.587
Nominalizations	-0.583
Academic words	-0.514
Incidence of s-bars	0.472

Incidence of singular nouns	-0.461
Minimum CELEX frequency sentence	0.411
Length of noun phrases	-0.408
Social words	0.400

Component 2: Text Brevity and Common N-Grams. The second component (see Table 3) appeared to represent text brevity and common n-gram use. From a brevity perspective, the component loaded shorter texts, texts with fewer word types, and text with fewer sentences. From an n-gram standpoint, the component loaded more frequent trigrams and bigrams for both written and spoken texts.

A correlation between the weighted scores for this component for each essay and the human ratings for essay quality reported $r(997) = -.554$, $p < .001$, indicative of a strong negative relation with essay quality. The correlation indicates that essays containing fewer words and more frequent n-grams are scored lower by expert raters.

Table 3: *Component 2 indices and loadings: Text brevity and common n-grams*

Index	**Eigen loading**
Written bigram frequency logarithm	0.880
Number of words	-0.855
Spoken trigram frequency logarithm	0.816
Written bigram frequency	0.803
Type count	-0.802
Number of sentences	-0.793
Written trigram frequency	0.763
Spoken trigram frequency	0.663

Component 3: Semantic Repetition. The third component (see Table 4) encapsulated semantic repetition, including higher text givenness, semantic similarity (i.e., between sentences and paragraphs), and referential overlap (i.e., word, stem, and argument overlap), as well as lower lexical diversity.

The correlation between the weighted scores for this component for each essay and the human ratings for essay quality, $r(997) = -.239$, $p < .001$, indicated a weak negative relation with essay quality. The correlation indicates that essays with greater repetition of words are of lower quality.

Table 4: *Component 3 indices and loadings: Semantic repetition*

Index	**Eigen loading**
LSA givenness	0.849
LSA sentence to sentence	0.824
LSA paragraph to paragraph	0.806

Content word overlap	0.715
Type token ratio	-0.709
Stem overlap	0.680
Lexical diversity (MTLD)	-0.662
Argument overlap	0.655
Lexical diversity (D)	-0.624

Component 4: Verbal Properties. The fourth component (see Table 5) captured verbal properties. For instance, this component loaded more verb phrases, incidence of infinitives, incidence of simple sentences (which contain a single main verb), and more verb base forms. The component also negatively loaded prepositions, prepositional phrases, and determiners, which are syntactically linked to nouns and not verbs.

The correlation between the weighted scores for this component for each essay and the human ratings for essay quality reported $r(997) = -.314$, $p < .001$, indicating a moderate negative relation with essay quality. The correlation demonstrates that essay with more verbs are of a lower quality.

Table 5: *Component 4 indices and loadings: Verbal properties*

Index	**Eigen loading**
Incidence verb phrases	0.794
Density verb phrases	0.794
Incidence of infinitives	0.713
Incidence of simple sentences	0.664
All verb incidence	0.597
Incidence of preposition phrases	-0.570
Incidence of prepositions	-0.554
Incidence of verb base forms	0.529
Incidence of determiners	-0.442

Component 5: Prototypical Words. The fifth component (see Table 6) represented prototypical words. The indices that loaded positively into this component included human ratings of concreteness (how concrete a word is), imageability (how imageable a word is), and meaningfulness (the number of associations a word contains).

The correlation between the weighted scores for this component for each essay and the human ratings for essay quality reported $r(997) = .188$, $p < .001$, indicative of a weak positive relation with essay quality. The correlations indicate that essays that contain more prototypical words (i.e., those that are more concrete, imageable, and have more associations) are scored higher by expert raters.

Table 6: *Component 5 indices and loadings: Prototypical words*

Index	Eigen loading
Word imageability	0.807
Word concreteness	0.751
Word meaningfulness	0.717

Component 6: Syntactic Simplicity. The sixth component (see Table 7) appeared to encapsulate syntactic simplicity. The component positively loaded syntactically similarity between sentences and paragraphs at the lexical and phrase level while negatively loading sentence length and number of words before the main verb, which are both indicators of syntactic complexity. The correlation between the weighted scores for this component and human ratings for essay quality, $r(997) = .008$, $p > .050$, indicated a negligible relation.

Table 7: *Component 6 indices and loadings: Syntactic simplicity*

Index	Eigen loading
Syntactic similarity across sentences	0.853
Average sentence length	-0.830
Syntactic similarity across paragraphs	0.825
Words before main verb	-0.365

Component 7: Future Time. The seventh component (see Table 8) seemed to capture future time. The component positively loaded more future words, modal verbs, and discrepancy words, which include modals and words such as *hope*, *desire*, and *expect*. The component negatively loaded a verb cohesion index, likely as a result of a lack of association between verb cohesion and auxiliary verb use. The correlation between the component scores and essay quality, $r(997) = -.217$, $p < .050$, reflected a weak relation with essay quality. Thus, essays that contain more future words are scored lower.

Table 8: *Component 7 indices and loadings: Future time*

Index	Eigen loading
Future words	0.806
Modal verbs	0.789
Discrepancy words	0.623
Verb cohesion	-0.482

Component 8: Nominal Simplicity. The eighth component (see Table 9) represented nominal simplicity. The component positively loaded more noun phrases, but negatively loaded adjectives, which can provide complexity to noun phrases as well as modify nouns. The correlation between the weighted scores for this component and the human ratings for essay quality, $r(997) = -.163$, $p < .050$, reflected a weak relation with essay quality. The correlations indicate that essays that contain more noun phrases without accompanying adjectives are scored lower.

Table 9: *Component 8 indices and loadings: Nominal simplicity*

Index	Eigen loading
Density noun phrases	-0.756
Incidence of adjectives	0.707
Incidence of noun phrases	-0.643
Incidence of adjectival phrases	0.624

Component 9: Global Cohesion. The ninth component (see Table 10) represented global cohesion. The component positively loaded indices that assessed links across large chunks of texts (paragraph types and essay to prompt). The correlation between the weighted scores for this component and the human ratings for essay quality, $r(997) = .113$, $p < .050$, reflected a weak relation with essay quality. The correlation demonstrates that essays with greater global cohesion were scored higher.

Table 10: *Component 9 indices and loadings: Global cohesion*

Index	Eigen loading
LSA body to body	0.789
LSA body to conclusion	0.787
LSA introduction to conclusion	0.657
LSA introduction to body	0.633
LSA essay to prompt	0.561

Regression Analysis: Component Scores Assumptions. None of the component scores were highly correlated with one another and, thus, all component scores that reported a significant correlation with essay quality were used (see Table 11 for correlations).

Table 11: *Correlations between component scores and essay quality*

Component score	r	p
Text brevity and common n-grams	-0.554	< .001
Lexical and nominal simplicity	-0.359	< .001
Verbal properties	-0.314	< .001
Semantic repetition	-0.239	< .001

Future time	-0.217	< .001
Prototypical words	-0.217	< .001
Nominal simplicity	0.188	< .001
Global cohesion	0.113	< .001
Syntactic simplicity	0.008	0.79

Training set. A regression analysis using the eight significant component scores as the independent variables to predict the human scores yielded a significant model, F (3, 670) = 155.718, p < .001, r = .641, R^2 = .411, for the training set.[4] Three component scores were included as significant predictors of the essay scores: *text brevity and common n-grams*, *lexical and nominal simplicity*, and *global cohesion*.

The model demonstrated that the component scores together explained 41% of the variance in the evaluation of the 673 independent essays in the training set (see Table 12 for additional information).

Table 12: *Stepwise regression analysis for component scores predicting essay quality*

Entry	**Component score added**	***r***	***R²***	***B***	**B**	***S.E.***
Entry 1	Text brevity and common n-grams	0.549	0.301	-0.003	-0.516	0
Entry 2	Lexical and nominal simplicity	0.636	0.405	-0.005	-0.314	0.001
Entry 3	Global cohesion	0.641	0.411	8.073	0.078	3.111

Note: *B* = unstandardized B B = standardized; S.E. = standard error. Estimated constant term is 2.668.

Test set. We used the model reported for the training set to predict the human scores in the test set. To determine the predictive power of the three component scores retained in the regression model, we computed an estimated score for each essay in the independent test set using the B weights and the constant from the training set regression analysis. This computation gave us a score estimate for the essays in the test set. A Pearson correlation was then calculated between the estimated score and the actual score assigned on each of the integrated essays in the test set. This correlation and the R^2 were then calculated to determine the predictive accuracy of the training set regression model on the independent data set.

When applied to the test set, the regression model reported r = .635, R^2 = .403. The results from the test set model demonstrated that the combination of the three component scores accounted for 40% of the variance in the assigned scores of the 304 essays in the test set, providing increased confidence for the generalizability of our model.

Exact and adjacent matches. We used the scores derived from the regression model to assess the exact and adjacent accuracy of the regression scores when compared to the human-assigned scores. For this analysis, we rounded the essay scores to the closest integer (i.e., a score of 4.5 was rounded to a 5). The regression model produced exact matches between the predicted essay scores and the human scores for 509 of the 977 essays (51% exact accuracy, see Table 13). The model produced exact or adjacent matches for 946 of the 977 essays (95% exact/adjacent accuracy). The measure of agreement between the actual score and the predicted score produced a weighted Cohen's Kappa of .515, demonstrating a moderate agreement.

Table 13: Confusion matrix for exact matches: Component analysis

Scores	1 (HS)	2 (HS)	3 (HS)	4 (HS)	5 (HS)
1 (MS)	**7**	4	0	0	0
2 (MS)	24	**58**	46	5	0
3 (MS)	11	91	**345**	173	30
4 (MS)	0	5	43	**96**	53
5 (MS)	0	0	0	3	**3**

Note: HS = human score; MS = machine score

Regression Analysis: Micro-features

Assumptions. We used the same indices that informed the component score analysis in the micro-feature regression analysis. Thus, none of micro-features demonstrated strong multicollinearity and all micro-features were normally distributed. All of the micro-features that reported a significant correlation with essay quality and reported at least a small effect size (i.e., $r > .100$) were included in the regression analysis (n = 68).

Training set. A regression analysis using the 68 significant micro-features as the independent variables to predict the human scores yielded a significant model, $F(10, 665) = 90.638$, $p < .001$, $r = .744$, $R^2 = .554$, for the training set. Ten micro-features were included as significant predictors of the essay scores: *type count*, *incidence of nominalizations*, *average paragraph length*, *type-token ratio*, *content word frequency*, *key type count*, *number of sentences*, *verb incidence*, *lexical diversity (D)*, and *verb hypernymy*.

The model demonstrated that the 10 micro-features explained 55% of the variance in the evaluation of the 675 independent essays in the training set (see Table 14 for additional information).

Table 14: *Stepwise regression analysis for micro-features predicting essay quality*

Entry	Micro-feature added	*r*	R^2	*B*	B	*S.E.*
Entry 1	Type count	0.560	0.313	0.006	0.262	0.001
Entry 2	Incidence of nominalizations	0.629	0.395	4.972	0.114	1.345
Entry 3	Average paragraph length	0.664	0.441	0.441	-0.229	0.006
Entry 4	Type-token ratio	0.691	0.478	-2.606	-0.242	0.408
Entry 5	Content word frequency	0.713	0.509	-0.888	-0.171	-0.187
Entry 6	Key type count	0.728	0.530	0.035	0.161	0.006
Entry 7	Number of sentences	0.734	0.539	0.024	0.194	0.006
Entry 8	Incidence of verbs	0.739	0.546	-0.004	-0.082	0.001
Entry 9	Lexical diversity (D)	0.742	0.551	0.005	0.105	0.002
Entry 10	Verb hypernymy	0.744	0.554	0.368	0.064	0.162

Notes: *B* = unstandardized B; B = standardized; S.E. = standard error. Estimated constant term is 4.699.

Test set. The regression model, when applied to the test set, reported r = .722, R^2= .522. The results from the test set model demonstrated that the combination of the 10 micro-features accounted for 52% of the variance in the assigned scores of the 302 essays in the test set, providing increased confidence for the generalizability of our model.

Exact and adjacent matches. The regression model produced exact matches between the predicted essay scores and the human scores for 539 of the 977 essays (54% exact accuracy, see Table 15). The model produced exact or adjacent matches for 958 of the 977 essays (96% exact/adjacent accuracy). The measure of agreement between the actual score and the predicted score produced a weighted Cohen's Kappa of .603, demonstrating a substantial agreement.

Table 15: Confusion matrix for exact matches: Microfeature analysis

Scores	**1 (HS)**	**2 (HS)**	**3 (HS)**	**4 (HS)**	**5 (HS)**
1 (MS)	**19**	2	0	0	0
2 (MS)	22	**95**	65	6	1
3 (MS)	1	56	**316**	164	26
4 (MS)	0	5	53	**106**	56
5 (MS)	0	0	0	1	**3**

Note: HS = human score; MS = machine score

Regression Analysis: Combined Micro-features and Component Scores

Assumptions. We used both the micro-features and the component scores that showed significant correlations and at least a small effect size with the essay scores for this analysis. Lexical Diversity D was strongly collinear with Component 3. Because Lexical Diversity D demonstrated a stronger correlation with essay quality, it was retained and the component score was removed. Number of words and number of types were strongly collinear with Component 2, but were removed because Component 2 showed a stronger relationship with essay quality. A similar result was found with the incidence of nouns and Component 1 leading to the removal of the incidence of nouns index. In total, this left us with 71 variables for the regression analysis.

Training set. A regression analysis using the 71 variables scores as the independent variables to predict the human scores yielded a significant model, F (8, 669) = 93.691, p < .001, r = .727, R^2 = .528, for the training set. Eight variables were included as significant predictors of the essay scores: Component 2 (Text Brevity and Common N-Grams), Component 1 (Lexical and Nominal Simplicity), number of paragraphs, key type count, word familiarity, incidence of nominalizations, incidence of verbs, and noun hypernymy.

The model demonstrated that the mixture of component scores and micro-features explained 53% of the variance in the evaluation of the 676 independent essays in the training set (see Table 16 for additional information).

Table 16: *Stepwise regression analysis for component scores and micro-features predicting essay quality*

Entry	**Index**	***r***	***R^2***	***B***	**B**	***S.E.***
Entry 1	Component 2 (Text brevity and common n-grams)	0.548	0.300	-0.002	-0.357	0
Entry 2	Component 1 (Lexical and nominal simplicity)	0.640	0.410	-0.001	-0.062	0.001
Entry 3	Number of paragraphs	0.676	0.456	0.155	0.239	0.020

Entry 4	Key type count	0.695	0.483	0.034	0.155	0.007
Entry 5	Word familiarity	0.707	0.500	-0.016	-0.114	0.005
Entry 6	Incidence of nominalizations	0.714	0.510	6.005	0.139	1.403
Entry 7	Incidence of verbs	0.722	0.521	-0.005	-0.113	0.001
Entry 8	Noun hypernymy	0.727	0.528	-0.165	0.090	0.050

Notes: *B* = unstandardized B B = standardized; S.E. = standard error. Estimated constant term is 2.668.

Test set. The regression model, when applied to the test set, reported r = .727, R^2 = .528. The results from the test set model demonstrated that the combination of the component scores and the micro-features accounted for 53% of the variance in the assigned scores of the 301 essays in the test set, providing increased confidence for the generalizability of our model.

Exact and adjacent matches. The regression model produced exact matches between the predicted essay scores and the human scores for 540 of the 977 essays (54% exact accuracy, see Table 17). The model produced exact or adjacent matches for 964 of the 977 essays (97% exact/adjacent accuracy). The measure of agreement between the actual score and the predicted score produced a weighted Cohen's Kappa of .614, demonstrating a substantial agreement.

Table 17: Confusion matrix for exact matches: Combined analysis

Scores	**1 (HS)**	**2 (HS)**	**3 (HS)**	**4 (HS)**	**5 (HS)**
1 (MS)	**18**	4	0	0	0
2 (MS)	21	**86**	56	3	0
3 (MS)	3	62	**323**	164	21
4 (MS)	0	6	55	**109**	61
5 (MS)	0	0	0	1	**4**

Note: HS = human score; MS = machine score

Comparison Between Regression Models

We used Fisher *r*-to-*z* transformation to assess the significance of the differences between the correlations reported for each regression model (i.e., the component score, micro-feature, and combined regression models). The z transformations demonstrated that both the micro-feature regression and the combined regression were significantly different than the component score regression. No differences were reported between the micro-feature and the combined regression (see Table 18 for details).

Table 18: *Fisher r-z transformation results between models*

Comparison	***z***	***p***
Micro-features > Component Scores	-1.98	< .050
Combined > Component scores	-2.11	< .050
Micro-features = Combined	-0.13	> .050

Discussion

This study examined the strengths of linguistic micro-features, component scores, and a combination of both in predicting essay quality. The findings support the notion that scoring models based on micro-features and a combination of micro-features and component scores are significantly better than scoring models based on component alone. This finding has important implications for AES and AWE systems and has strong potential to be used to guide the development of both. The study is also unique because it introduces a statistically viable method to develop aggregated scores (i.e., a principal component analysis) and provides explicit information about the micro-features used and how these micro-features are combined into aggregated features. This information should allow for more precise replications and afford researchers and teachers with specific information about how micro-features are computed and how they can be combined to assess larger linguistic constructs. Below, we discuss the developed component scores, the scoring models, the accuracy of the scoring models, the comparison of these scoring models, and how these models may aid in formative feedback to users in AWE systems.

Our factor analysis led to the development of nine aggregated features, eight of which demonstrated significant correlations with essay quality. Of these eight components, the strongest predictor of essay quality was related to text length and n-gram production such that higher-quality essays were longer and contained less common n-grams. Four components were related to the production of nouns and verbs. The correlations for these features indicated that essays that contained fewer simplistic noun phrases, more sophisticated words, and more verbal properties were scored higher. The exception was the use of future time, which was related to lower essay quality. Two components were related to cohesion with the first related to semantic repetition and the second related to global cohesion (i.e., links between larger chunks of text such as paragraphs). The semantic repetition indices correlated negatively with essay quality, indicating that less semantic repetition is indicative of more successful writing. The global cohesion indices correlated positively, suggesting that essays that linked paragraphs together were more successful. Only one component score did not correspond to human judgments of writing quality: syntactic complexity. Overall, the component scores indicated essays that were longer, had greater verbal properties, contained more complex noun phrases, were more lexically sophisticated (both with words and n-grams), and had greater global cohesion but less semantic repetition were scored higher. This finding is in line, to some degree, with previous aggregated scores (see Attali & Burstein, 2006); however, the component scores reported here are more specific to the tools used and thus more linguistic in nature. We did not develop aggregated scores related to usage, grammar, and mechanics, but did develop aggregated scores related to verbal and nominal properties, prototypical words, different types of cohesion, and future time.

A combination of three of these component scores (*Text brevity and common n-grams*, *Lexical and nominal simplicity*, and *Global cohesion*) reported a correlation of $r = .641$ and explained 41% of the variance in the training set and 40% of the variance in the test set. Overall, the model based on these features had an exact accuracy of 51% and an exact/adjacent accuracy of 95%. This is generally equivalent to that reported by other AES systems, but both of these accuracies are a bit lower than expected. The *r* values and reported Kappa values were also lower than that reported for the human raters. This may be because the model is matched to average scores between two raters as compared to one rater (as found in other scoring models) or because the tools used in this analysis did not include component scores related to grammar, usage, style, or development, which are common in other AES systems such as e-rater (Enright & Quinlan, 2010) and Intellimetric (Dikli, 2006; Schultz, 2013). However, such concerns should also be an issue with the micro-feature and mixed regression models, but the results from these models were stronger.

For instance, our regression analysis using micro-features yielded a significant model that reported a correlation of $r = .744$, and explained 55% of the variance in the training set and 52% of the variance in the test set using nine variables related to text length (number of paragraphs and sentences), nouns, verbal properties, lexical sophistication, breadth of vocabulary, key word use, and cohesion. Overall, the model based on these features had an exact accuracy of 54% and an exact/adjacent accuracy of 96%. This is more in line with the accuracies reported by other AES systems when compared to the component score regression and is in line with *r* and Kappa values for the human ratings. In addition, the Fisher *r*-to-*z* transformations demonstrated that the correlation reported for the micro-feature regression was significantly better than that reported for the component score regression. In total, the findings from this model show that essay quality is related to the use of a greater number of word types, a greater number of sophisticated words (more nominalizations, more verbs with a greater number of senses, and more infrequent words), greater text length (more paragraphs and sentences), less word repetition (type token ratio), more key word types, and more verbs.

Our regression analysis using both component scores and micro-features reported a significant model. The model yielded a correlation of $r = .727$ and explained 53% of the variance in the training set. The model relied on eight features, two of which were

component scores (*text brevity and common n-grams* and *Lexical and nominal simplicity*) and six of which were micro-features (number of paragraphs, key type count, word familiarity, nominalizations, incidence of verbs, and noun hypernymy). Overall, the model based on these features had an exact accuracy of 54% and an exact/adjacent accuracy of 97%. Like the micro-feature analysis, this is similar to that reported by other AES systems and by the human raters in this study. Also, like the micro-feature analysis, the Fisher r-to-z transformations demonstrated that the correlation reported for the mixed regression was significantly better than that reported for the component score regression. However, no differences were reported between the mixed regression and the micro-feature regression. Overall, the findings from the mixed model show that essay quality is related to greater text length, nominal simplicity, lexical sophistication (both n-grams and single words), number of paragraphs, the production of key types, and the use of verbs.

In total, these analyses indicate that AES models that use either micro-features alone or a mix of micro-features and component scores out-perform AES models that use component scores alone. Such a conclusion is based on the correlations reported by the regression models and the exact and exact/adjacent accuracies reported. In all cases, the accuracies reported by the micro-features and the mixed models were superior to those reported by the component scores only model. In the case of the correlations, the differences were significant. In the case of the exact and the exact/adjacent accuracies, the reported Kappas for the micro-features and mixed models reached the level of substantial agreement whereas the Kappa for the aggregated model reached the level of moderate agreement.

Overall, the three models depended on similar features. The shared features among the models included text length, lexical sophistication, and nominal simplicity with text length always the strongest feature. The strength of text length in the regression model may explain why exact matches reported for all models are higher for lower scored essays than higher scored essays (see Tables 13, 15, and 17) in that essay length is a strong predictor of lower scored essays but not higher scored essays (Crossley et al., 2014). Missing from the micro-feature and mixed analysis but included in the component score analysis were indices related to global cohesion. However, it should be noted that global cohesion accounted for only 1% of the variance in the component score regression. In addition, the micro-feature analysis included two indices of cohesion (lexical repetition) that explained about 3% of the variance. Missing from the component score regression, but shared in the micro-feature and mixed models were variables related to key word use, and paragraph length. These variables explained 6% and 10% of the variance in the micro-feature and mixed models respectively.

In general, a greater number and variety of indices informed the micro-feature and the mixed models. An increased number of features should theoretically provide the opportunity to give a greater depth and breadth of feedback to users of an AWE system. With the component score model reported in this study, we could only provide users with feedback on text length, n-gram use, lexical and nominal simplicity, and global cohesion. Using our micro-feature model, we would not be able to provide feedback on global cohesion, but we could provide additional feedback in terms of lexical breadth (i.e., the number of word types used), paragraph and sentence length, lexical repetition (cohesion), key type use, and verbal properties. Our mixed model would not provide feedback on cohesion; but, like the micro-feature model, it could provide additional feedback in terms of number of paragraphs, key type use, and verbal properties. Thus, not only are the micro-feature and mixed models more accurate in providing summative feedback (i.e., matching holistic scores), they should also afford greater opportunities to provide relevant feedback to AWE users.

It is also worth noting that the nine component scores only explained 37% of the variance within the essays. While this variance is not linked to estimated quality, it does indicate that there are numerous factors beyond the linguistic variables calculated for this study that interact with essay writing. In some respects, this is expected because there was not direct overlap between the developed components and the all the language features found in the rubric. While there was overlap between many of the linguistic features in the rubric and the component scores (i.e., the component scores covered language features such as vocabulary, cohesion, appropriate examples, and sentence variety), there were a number of features not covered by the component scores. These included linguistic features such as grammar and spelling, but also features related to point of view and critical thinking. It is also likely that variables not found in the rubric or the component scores such as argument strength, prompt, and writer attributes (i.e., individual differences such as age, gender, aptitude, literacy skills) may explain additional variance within the human scores of essay quality.

Conclusion

This study introduces a novel approach to developing aggregated scores and compares this approach to using micro-features alone and a combination of micro-features and aggregated scores. Overall, the findings from this study support the use of micro-features or a mix of micro-features and component scores in developing scoring models for AES systems. Models developed using these features show significant gains over those using component scores alone and the number and variety of features available in these models provide the possibility of giving a greater depth of feedback to users in an AWE system. We find no strong evidence that a micro-feature approach leads to models of essay quality that contain redundant variables nor do we find that the reported models are not stable, as seen in the similarity between or training and test set results. Lastly, while the component approach leads to more

representative constructs, both the micro-feature and mixed models included similar variables along with a greater number of divergent variables.

It should be noted, however, that the findings from this study are specific to the writing prompts, tasks, and the tools used. That is to say, these findings may not transfer to other prompts or writing tasks, although they do appear generalizable across a number of independent writing prompts. In reference to the tools used, Coh-Metrix, LIWC, and WAT do not calculate micro-features that examine grammar or mechanics, which are common in other AES systems.[5] In addition, the tools used in this study may not measure elements of style or development in the same manner as found in other AES systems such as e-rater and Intellimetric. Thus, the use of different tools to calculate different variables may lead to more or less robust findings then reported here for aggregated scores. Future studies should consider such possibilities along with using larger sets of scored essays if available.

Notes

1. AWE systems are similar to AES systems but, in addition to providing a predicted essay quality score, they also give feedback on student writing.

2. Attali and Powers (2008) used a confirmatory factor analysis to determine the construct coverage of e-rater micro-features in examine if there were similar underlying structures for essays written across grade levels (in this case 4th grade to 12th grade). They reported a three-model solution corresponding to “fluency” (essay length and style features), conventions (grammar usage and mechanics), and word choice (vocabulary and word length) that defined the structure of essays written at various grade levels. Somasundaran et al. (2014) used a PCA to reduce features sets related to discourse coherence.

3. A weighted Kappa takes into account the degree of disagreement between observed scores.

4. Component six, Syntactic Simplicity, was not included in the regression model because it was not significantly correlated with essay score.

5. In the case of Coh-Metrix and WAT, grammar and mechanics indices are not calculated because a number of studies demonstrate that these features are not strong predictors of essay quality (Crossley et al., 2014), nor does feedback on these features lead to improved essay quality (Graham & Perin, 2007).

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. We thank our expert raters for their assistance in scoring the essays used in this study.

Bios

Scott Crossley is an Associate Professor at Georgia State University. His interests include computational linguistics, corpus linguistics, and second language acquisition. He has published articles in second lexical acquisition, second language writing, second language reading, discourse processing, language assessment, intelligent tutoring systems, and text linguistics.

Kris Kyle is a PhD candidate at Georgia State University in the Department of Applied Linguistics and ESL. His research interests include corpus linguistics, second language writing, computational linguistics, and second language assessment.

Danielle S. McNamara is a Professor in the Psychology Department at Arizona State University. She focuses on educational technologies and discovering new methods to improve students’ ability to understand challenging text, learn new information, and convey their thoughts and ideas in writing. Her work integrates various approaches and methodologies including the development of game-based, intelligent tutoring systems (e.g., iSTART, Writing Pal), the development of natural language processing tools (e.g., iSTART, Writing Pal, Coh-Metrix, the Writing Assessment Tool), basic research to better understand cognitive and motivational processes involved in comprehension and writing, and the use of learning analytics across multiple contexts. More information about her research and access to her publications are available at soletlab.com.

References

Attali, Y. (2008). *E-rater performance for TOEFL iBT independent essays*. Unpublished manuscript.

Attali, Y., & Powers, D. (2008). A developmental writing scale. *ETS Research Report Series, 2008*(1). Princeton, NJ: ETS.

Aull, L. L., & Lancaster, Z. (2014). Linguistic markers of stance in early and advanced academic writing: A corpus-based

comparison. *Written Communication, 31*(2), 151-183.

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre?. *Reading and Writing, 22*(2), 185-200.

Bereiter, C. (2003). Foreword. In M.D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary approach* (pp. vii-ix). Mahwah, NJ: Lawrence Erlbaum Associates.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Burstein, J. (2003). The e-rater scoring engine: Automated Essay Scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary approach* (pp. 113-121). Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online service. *AI Magazine, 25*, 27-36.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In M.D. Shermis and J. Burstein (Eds.). *Handbook of automated essay evaluation: Current applications and new directions*(pp. 55-67). Routledge, New York and London.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*, 100-108.

Connor, U. (1990). Linguistic/rhetorical measures of international persuasive student writing. *Research in the Teaching of English, 24*, 67-87.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A*, 497-505. doi: 10.1080/14640748108400805

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D.S. (2014). Analyzing Discourse Processing Using the Simple Natural Language Processing Tool (SiNLP). *Discourse Processes, 51* (5-6), 511-534.

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In P. M. McCarthy & G. M. Youngblood (Eds.). *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 214-219). Menlo Park, CA: The AAAI Press.

Crossley, S. A., Kyle, K., Allen, L, & McNamara, D. S. (2014). How important are grammar and mechanics in writing assessment and instruction? Evidence from essay analyses. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.). Proceedings of the 7th Educational Data Mining (EDM) Conference (pp. 300-303). Heidelberg, Germany: Springer.

Crossley, S.A., & McNamara, D.S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 53*, 115-136.

Crossley, S. A., Roscoe, R. D., & McNamara, D. S. (2013). Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. In P. M. McCarthy & G. M. Youngblood (Eds.), *Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 208-213). Menlo Park, CA: The AAAI Press.

Crossley, S. A., Roscoe, R. D., McNamara, D. S., & Graesser, A. (2011) Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education*. (pp. 438-440). New York: Springer.

Crossley, S. A., Weston, J., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*(3), 282-311.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7-24.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment, 5*(1).

Elbow, P., & Belanoff, P. (1986). Portfolios as a substitute for proficiency examinations. *College Composition and Communication*, 336-339.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater(R) scoring. *Language Testing, 27*(3), 317-334.

Fellbaum, C. (1998). *WordNet*. Cambridge, MA: MIT Press.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414-420.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M.D. Shermis & J. Burstein (Eds.). *Handbook of automated essay evaluation: Current applications and new directions* (pp. 68-88). Routledge: New York and London.

Frase, L., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL test of written English*. (TOEFL Research Report No. 64). Princeton, NJ: ETS.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Writing Assessment, 18*, 218-238.

Graesser, A.C., McNamara, D.S., & Kulikowich, J.M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223-234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers, 36*, 193-202.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*(3), 445-476.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*(2), 123-145.

Haswell, R. H. (2006). Automatons and automated scoring: Drudges, black boxes, and dei ex machina. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 57-78). Logan, UT: Utah State University Press.

Haswell, R., & Ericsson, P. (Eds.). (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Hempelmann, C. F., Dufty, D., McCarthy, P. M., Graesser, A. C., Cai, Z., & McNamara, D. S. (2005). Using LSA to automatically identify givenness and newness of noun phrases in written discourse. In *Proceedings of the 27th annual conference of the Cognitive Science Society* (pp. 941-946). Retrieved from

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

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language, 25*(2), 282-306.

Huot, B. (1996). Towards a new theory of writing assessment. *College Composition and Communication, 47*, 549-566.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57-84.

Jin, W. (2001). *A quantitative study of cohesion in Chinese graduate students' writing: Variations across genres and proficiency*

levels (ERIC Document Reproduction Service No. ED 452 726).

Landauer, T. K., Laham, R. D. & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M.D. Shermis & J.C. Bernstein (Eds.), *Automated Essay Scoring: A cross-disciplinary perspective*(pp. 87-112). Mahwah, NJ: Lawrence Erlbaum Publishers.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ: Erlbaum.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *LSA: A road to meaning*. Mahwah, NI: Lawrence Erlbaum Associates.

Malvern, D. D. Richards, B. J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan. doi: 10.1057/9780230511804

Mayfield, E., & Ros é, C. P. (2013). LightSIDE: Open source machine learning for text. In M.D. Shermis and J. Burstein (Eds.). *Handbook of automated essay evaluation: Current applications and new directions* (pp. 124-135). Routledge: New York and London.

McCarthy, P.M. & Jarvis, S., (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 381-392.

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P.M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188-205). Hershey, PA: IGI Global.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication, 27*(1), 57-86.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods, 45*(2), 499-515.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

McNamara, D. S., Raine, R., Roscoe, R., Crossley, S., Jackson, G. T., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Lam, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P. M., & Graesser, A. C. (2012). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 298-311). Hershey, PA: IGI Global.

Myers, M. (2003). What can computers and AES contribute to a K-12 writing program? In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 3-20). Mahwah, NJ: Lawrence Erlbaum Associates.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Mahwah, NJ: Lawrence Erlbaum Associates.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *LIWC2007: Linguistic inquiry and word count*. Austin, Texas.

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 121-131). Fort Collins, Colorado: WAC Clearinghouse/Anderson, SC: Parlor Press.

Perelman, L. (2014). When “the state of the art” is counting words. *Assessing Writing, 25*(3), 104-111. doi: 10.1016/j.asw.2014.05.001

Ramineni, C., Trapani, C. S., Williamson, D. M. W., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater ® scoring engine*

for the TOEFL ® independent and integrated prompts. (ETS Research Report No. RR-12-06). Princeton, NJ: ETS.

Reid, J. (1986). Using the Writer's Workbench in composition teaching and testing. In C. Stansfield (Ed.), *Technology and language testing* (pp. 167-188). Alexandria, VA: TESOL.

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191-210). Cambridge: Cambridge University Press.

Reppen, R. (1994). A genre-based approach to content writing instruction. *TESOL Journal, 4*(2), 32-35.

Roscoe, R., Kugler, D., Crossley, S. A., Weston, J., & McNamara, D. S. (2012). Developing pedagogically-guided threshold algorithms for intelligent automated essay feedback. In P. M. McCarthy & G. M. Youngblood (Eds.). *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 466-471). Menlo Park, CA: The AAAI Press.

Rudner, L., Garcia, V., & Welch, C. (2005). *An evaluation of Intellimetric ™ essay scoring system using responses to GMAT ® AWA prompts* (GMAC Research report number RR-05-08).

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 from

Schultz, M. T. (2013). The IntelliMetric automated essay scoring engine: A review and an application to Chinese essay scoring. In M.D. Shermis and J. Burstein (Eds.). *Handbook of automated essay evaluation: Current applications and new directions* (pp. 89-98). Routledge: New York and London.

Shermis, M.D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw, & N. S. Petersen (Eds.), *International encyclopedia of education* (3rd Ed.) (pp. 20-26). Oxford, UK: Elsevier.

Shermis, M. D., & Hamner, B. (2013). 19 contrasting state-of-the-art automated scoring of essays. In M.D. Shermis & J. Burstein (Eds.). *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313-346). Routledge: New York and London.

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

Wardle, E., & Roozen, K. (2013). Addressing the complexity of writing development: Toward an ecological model of assessment. *Assessing Writing, 17*, 106-119.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*(1), 22-36.

Warschauer, M. & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 157-180.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335-353.

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 1*, 85-99.

Witten, I. H., Frank, E., Hall, M. A., (2011). *Data mining: Practical machine learning tools and techniques* (3rd Ed.). Morgan Kaufmann, San Francisco.

Witte, S. P., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College composition and communication, 32*(2), 189-204.

Yang, W. W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28*, 53-67.

Appendix : Example Prompts and Assignments

Fitting in Prompt and assignment
From the time people are very young, they are urged to get along with others, to try to "fit in." Indeed, people are often rewarded for being agreeable and obedient. But this approach is misguided because it promotes uniformity instead of encouraging people to be unique and different. Differences among people give each of us greater perspective and allow us to make better judgments.

Is it more valuable for people to fit in than to be unique and different? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.

Winning prompt and assignment
From talent contests to the Olympics to the Nobel and Pulitzer prizes, we constantly seek to reward those who are "number one." This emphasis on recognizing the winner creates the impression that other competitors, despite working hard and well, have lost. In many cases, however, the difference between the winner and the losers is slight. The wrong person may even be selected as the winner. Awards and prizes merely distract us from valuable qualities possessed by others besides the winners.

Do people place too much emphasis on winning? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.

Appendix B: Example Essays

Fitting in Essay (Human score = 2)
I honestly believe that the question sends people to two totally opposite ends of the spectrum and speaks of the same people. In my opinion neither person is different because they are both followers. People who try and fit in are following the crowd that goes one way and does the “popular” thing. People who try to be unique and different are followers as well because they follow a group or idea that being different is an independent idea but all of the people you see trying to be different have to fit into a group of “unique people.”

I personally feel like I am a strong leader. Coming from my personal experience you can fit in but still be different and unique. Although I may be influenced by ideas or clothing styles or many other things. Although I may fit in by my style or by my personality, I have an opportunity to be different and unique by my beliefs morals and faith. People who fit in but who are different from the people in that group are the ones who come to lead people and influence them most of the time in a good way. I think people who are in the crowd but don't follow the crowd are the ones who find the most value in this and usually those people can influence the "in crowd" as well as those who try to be unique and different. If you look at both categories of people, they usually both end up in the same place in life. They are constantly looking for something more or they are involved in crime. People who are followers usually tend to follow the wrong group of people. If you are a follower of God usually things produced from you are good but those who follow groups that dwell in materialism cause things such as what is going on with America today.

Winning (Human score = 5)
Yes, I do believe we place too much emphasis on winning. To most people, lie is some sort of contest that each person is trying to win. Each person wants the best job, the nicest car, the big house with a pool, and every other nice thing they can think of. Most people do not feel as if they have "won" this game of life unless they get all of these things that in reality they do not even need. I believe we should lessen the emphasis on winning because it takes away from those who have worked as hard as they possibly could and still lost.

One example that makes me think this is in the case of my girlfriend. She has participated in many singing contest and beauty pageants. She has won a few, and has also lost a few. Everything is fine and dandy when she wins, but when she does not get first place, it is very upsetting for her. I see first hand the work she puts in to each competition and how many hours she spends trying to perfect the art she is practicing. She is one of the hardest working people I know and when she does not get first, it seems that all that practice is for nothing. For instance, in one of the beauty pageants she lost, it was not even because the other girls were better. She was clearly one of the best out there, but because of her age, she was not chosen as the winner, or even runner up. The judges specifically told her that her age was a big reason for her not winning. Is this right? After countless hours of work and late nights without sleep because of practicing, for her to be good enough, but not win because of age? If there was not such an emphasis on winning first place, it would be easier for her to see that the experience was a great one for her, and that she did the absolute best she could. She would also have more confidence in her abilities. Instead, it makes her think that she is not as good when she really is.

Another example of the negative effects of placing too much emphasis on winning is with the Olympics. I personally know people who train everyday for the Olympics. This contest comes around every four years and is held in many different places around the

world and is between many different countries. One of my friends trains for gymnastics every day. She wakes up at 4 AM every morning to go to the gym to train, goes to school from 7 AM to around 2:30 PM, then goes straight to the gym from school and trains from 3 PM to around 9 PM. Not only does she train that much every weekday, but all day Saturday and most of the day Sunday is spent in training. That's nearly 60 or more hours a week spent training for events that will last for about a week. So, she trains this hard for years leading up to the Olympics. What happens if during one of her routines, she makes one tiny mistake or mental lapse and loses? In her mind, that one tiny mistake makes all of those hours practiced mean nothing. She does not mean to think this way, but because of the emphasis put on winning the gold medal, or getting in first place, she feels like she has failed if she does not win. Is it right for someone to work that hard and lose? I do not believe so. I believe that if someone gives everything they have to something, then they should be considered a winner, and that emphasis on first place or failure should be abolished.

These are just two examples of many I could give of why too much emphasis being placed on winning is a bad thing. There are so many people that devote their lives to something and end up being called a "loser". Now, I am not saying that winning is a bad thing. To those who devote their lives to something and win their dream, there is absolutely nothing wrong with that. However, for those who devote their lives to something and lose, it is not right for the world to make them feel like failures.

Copyright © 2021 - ***The Journal of Writing Assessment*** - All Rights Reserved.