Assessing Writing 23 (2015) 35–59

Contents lists available at ScienceDirect

Assessing Writing

A hierarchical classification approach to automated essay scoring

Danielle S. McNamara [a,*], Scott A. Crossley [b], Rod D. Roscoe [a], Laura K. Allen [a], Jianmin Dai [a]

[a] Arizona State University, United States
[b] Georgia State University, United States

ARTICLE INFO

Article history:
Received 22 September 2013
Received in revised form 2 September 2014
Accepted 4 September 2014

Keywords:
Automated essay scoring
AES
Writing assessment
Hierarchical classification

ABSTRACT

This study evaluates the use of a hierarchical classification approach to automated assessment of essays. Automated essay scoring (AES) generally relies on machine learning techniques that compute essay scores using a set of text variables. Unlike previous studies that rely on regression models, this study computes essay scores using a hierarchical approach, analogous to an incremental algorithm for hierarchical classification. The corpus in this study consists of 1243 argumentative (persuasive) essays written on 14 different prompts, across 3 different grade levels (9th grade, 11th grade, college freshman), and four different time limits for writing or *temporal* conditions (untimed essays and essays written in 10, 15, and 25 minute increments). The features included in the analysis are computed using the automated tools, Coh-Metrix, the Writing Assessment Tool (WAT), and Linguistic Inquiry and Word Count (LIWC). Overall, the models developed to score all the essays in the data set report 55% exact accuracy and 92% adjacent accuracy between the predicted essay scores and the human scores. The results indicate that this is a promising approach to AES that could provide more specific feedback to writers and may be relevant to other natural language computations, such as the scoring of short answers in comprehension or knowledge assessments.

* Corresponding author at: P.O. Box 872111, Tempe, AZ 85287-2111, United States. Tel.: +1 480 727 5690.
E-mail address: dsmcnamara1@gmail.com (D.S. McNamara).

http://dx.doi.org/10.1016/j.asw.2014.09.002
1075-2935/

1. Introduction

Teaching students how to write well is a fundamental objective of our educational system for obvious reasons. Students who cannot write well are less likely to effectively convey their ideas, persuade others, and succeed in various personal and academic endeavors. However, writing instruction takes an inordinate amount of teacher time, not only for the instruction of how to write but also in scoring essays and providing subsequent feedback to students. Done well, essay scoring is an enormously complex cognitive task that involves a multitude of inferences, choices, and preferences on the part of the grader. What features are attended to, which characteristics and sections are weighted most highly, and what standards are held are all factors that may vary widely across human graders. Indeed, essay ratings are highly variable from human to human (Huot, 1990, 1996; Meadows & Billington, 2005).

A solution to this variability across raters has been to train expert raters to use scoring rubrics (Bridgeman, 2013). For example, the SAT asks students to write essays in response to prompts such as those presented in Table 1. The SAT rubric for persuasive writing (College Board, 2011; see Appendix) includes six levels that address writers' critical thinking, use of examples and evidence, organization and coherence, language and vocabulary, sentence structure, and mechanics. For example, high scoring essays that receive a score of 6 are classified as using "clearly appropriate examples, reasons, and other evidence" and exhibiting "skillful use of language, using a varied, accurate, and apt vocabulary" whereas low-scoring essays receiving a score of 1, provide "little or no evidence" and display "fundamental errors in vocabulary." While the reliability of human scores using such rubrics (with training and examples) is quite high, essay scoring remains relatively time demanding, be it for a teacher tasked to score 150 essays over the weekend, or for a company challenged to score thousands of essays for the purpose of standardized assessment. The increased recognition of the importance of writing, combined with cost considerations and the obvious time demands to reliably and validly score writing, heightens the need for more rapid feedback and, by consequence, has fed the growth of research on automated essay scoring (AES; Dikli, 2006; Graesser & McNamara, 2012; Shermis & Burstein, 2013; Weigle, 2013; Xi, 2010).

The focus of this study is to describe a new method of AES that we have designed using hierarchical classification and report on its reliability in comparison to more common scoring models that have been reported in the literature. AES technologies have been largely successful, reporting levels of accuracy that are in many situations as accurate as expert human raters (Attali & Burstein, 2006; Burstein, 2003; Elliott, 2003; Landauer, Laham, & Foltz, 2003; Rudner, Garcia, & Welch, 2006; Shermis, Burstein, Higgins, & Zechner, 2010; Streeter, Psotka, Laham, & MacCuish, 2002; Valenti, Neri, & Cucchiarelli, 2003). AES systems assess essays using a combination of computational linguistics,

Table 1
SAT instructions and examples of SAT writing prompts and assignments.

SAT instructions	Your essay must be written on the lines provided on your answer sheet – you will receive no other paper on which to write. You will have enough space if you write on every line, avoid wide margins, and keep your handwriting to a reasonable size. Remember that people who are not familiar with your handwriting will read what you write. Try to write or print so that what you are writing is legible to those readers.
Prompt 1	Think carefully about the following statement. Then read the assignment below it and plan and write your essay as directed. "The more things change, the more they stay the same." Assignment: Do you agree with this statement? Plan and write an essay in which you develop your position on this issue. Support your point of view with reasoning and examples taken from your reading, studies, experience, or observations.
Prompt 2	Consider carefully the following statement. Then read the assignment below it and plan and write your essay as directed. "It is as difficult to start things as it is to finish things." Assignment: Do you agree with this statement? Plan and write an essay in which you develop your position on this issue. Support your point of view with reasoning and examples taken from your reading, studies, experience, or observations.

Note: Additional examples of SAT writing prompts are available from the following websites: http://www.bcps.org/offices/cte/pdf/SAT-Writing-Prompts.pdf, http://www.collegeboard.com/student/testing/sat/after/essay_prompts.html, http://www.sparknotes.com/testprep/books/newsat/powertactics/essay/chapter7.rhtml.

statistical modeling, and natural language processing (Shermis & Burstein, 2013). For example, systems such as e-rater developed at Educational Testing Service (Burstein, Chodorow, & Leacock, 2004; Burstein, Tetreault, & Madnani, 2013) and the IntelliMetric Essay Scoring System developed by Vantage Learning (Rudner et al., 2006; Schultz, 2013) rely primarily on combinations of natural language processing techniques and artificial intelligence, whereas the Intelligent Essay Assessor (Foltz, Streeter, Lochbaum, & Landauer, 2013; Landauer et al., 2003) primarily relies on Latent Semantic Analysis.

Across AES systems, a typical methodology is followed. First, a set of target essays are divided into a training set and a test (or validation) set. A computational algorithm is tuned to optimally fit the essays in the training set using features automatically calculated from the text. The quantitative solution for the training set is typically a linear multiple regression formula or a set of Bayesian conditional probabilities between text features and scores. Many AES systems are commercialized and thus the details of the models and the calculated text features are oftentimes not released. Nonetheless, for the most part, AES systems tend to rely on a combination of threshold and regression analysis techniques. That is, the text variables that are selected to predict the human score of the essay are regressed (in a broad sense) onto the score using statistical techniques, such as machine learning algorithms, linear regressions, or stepwise regressions. In some cases, thresholds are set such that the essays must reach a certain value to receive a particular score. The quantitative solution that results from the training set algorithm is then applied to a test set that has been set aside, and these scores are compared to the scores of the human raters.

The algorithm is considered successful if the scores from the algorithm and humans are relatively equivalent (Bridgeman, 2013). In terms of producing an automated score that closely matches a human score, these techniques work quite well. In one of our recent studies (McNamara, Crossley, & Roscoe, 2013), a combination of eight variables was able to account for 46% of the variance in human ratings of essay quality (the reported inter-rater reliability was $r > .75$). The predictions from this model resulted in perfect agreement (exact match of human and computer scores) of 44% and adjacent agreement (i.e., within 1 point of the human score) of 94% in a set of 313 essays. The weighted Cohen's kappa for the adjacent matches was .401, which demonstrates a moderate agreement. Across studies, human and computer-based scores correlate from .60 to .85, and several systems report perfect agreement from 30 to 60% and adjacent agreement from 85 to 100% (Attali & Burstein, 2006; Rudner et al., 2006; Shermis et al., 2010; Warschauer & Ware, 2006).

When the goal of a system is solely to provide a score, research and development are motivated primarily to increase accuracy by combining different types of linguistic, semantic, and rhetorical features of essays, and using different statistical and machine learning techniques (e.g., decision trees, Bayesian probabilities, regression). Overall, this approach is generally successful or acceptable. That is, automated scores tend to be similar to the scores a trained human would have assigned. If a student received a 4 on an essay from a human, an AES system would be highly likely to give the essay a 3, 4, or 5, with the highest probability of perfect agreement (i.e., a score of 4). The overarching goal is to match human scores, which generally fall on a 1 to 6 scale of some sort.

The above approach to AES has had two principal uses. First, it has been used by the assessment industry to facilitate the grading of essays (Dikli, 2006; Shermis & Burstein, 2003). Each year, millions of students author essays as part of high-stakes standardized testing (e.g., the Test of English as a Foreign Language and the Graduate Record Exam), and one motivation for AES has been to automate the scoring of such essays. These same technologies can be applied to help teachers grade lower-stakes writing assignments in class. Second, AES has been incorporated into instructional systems that allow students to write essays and receive automated feedback on their writing (Grimes & Warschauer, 2010; Warschauer & Grimes, 2008). In this case, the objective goes beyond solely providing an accurate score. The system may provide a score as a means of general feedback, like a grade, but also offers feedback regarding errors the student has made and ways to improve the essay through revision. In fact, as a means of separation, such systems are no longer referred to as AES systems, but rather automatic writing evaluation (AWE) systems.

In the literature, essay scoring has largely been the focus of research and development for AES systems. However, although the cutoffs to assign the score based on a regression analysis may be somewhat arbitrary, essay feedback can in some cases be directly tied to the scoring algorithm. For example, many algorithms and systems focus on the lower-level "traits" of the essays, such as the mechanics,

grammar, and spelling (Attali & Burstein, 2006; Burstein et al., 2004; Rudner et al., 2006; Shermis, Koch, Page, Keith, & Harrington, 2002). When the algorithm is based on the number of mechanical errors, grammatical mistakes, misspellings, the number of words in the essays, the number of paragraphs, and so on, then feedback can be facilely tied to those components of the algorithm (once the thresholds are set). For instance, if a given essay displays frequent sentence fragments, then feedback on appropriate punctuation and grammar may be given. Indeed, this appears to be the industry standard for the most part with many of the currently available systems focusing on providing feedback on lower-level traits (Kellogg, Whiteford, & Quinlan, 2010).

Unfortunately, recent research indicates that feedback on lower-level traits does little to improve essay performance. Graham and Perin's (2007) meta-analysis of writing interventions indicated that instruction focused on grammar and spelling was the least effective (Cohen's $d = -.32$), and perhaps deleterious type of feedback available, whereas the most effective interventions were those that explicitly and systematically provided students with instruction on how to use strategies for planning, drafting, editing, and summarizing (Cohen's $d = .82$). While these results should be interpreted with caution (see e.g., Fearn & Farnan, 2005), these findings nevertheless suggest that writing feedback should not be solely focused on the grammatical accuracy of individual essays. In particular, AWE systems may benefit from placing a greater emphasis on writing strategy instruction and tackling the challenge of providing feedback that addresses deeper aspects of the essay, particularly feedback that can point writers toward beneficial strategies. Of the current systems that are focused on feedback to the writer, few of them provide feedback on what strategies a student might use to improve the essay (e.g., Attali & Burstein, 2006). That is, the feedback may indicate what is weak within the essay at one level or another, but there are few systems that are able to follow the recommendation that can be inferred from Graham and Perin's (2007) meta-analysis indicating that feedback should point the writer toward strategies to improve the essay.

One level of this problem is pedagogical – what strategies should the writer be told to use and when? That issue is not the focus of this study. The other level of this problem (i.e., the focus of this study) is how to develop a computational algorithm that has some potential to afford the capability to link the algorithm output to strategy-focused feedback for the writer. There are several barriers to reaching this goal. First, as mentioned earlier, there is often little obvious connection between AES algorithms and writing strategies. If an algorithm were assessing primarily lower level mechanics, then it would be nearly impossible to tie the outcome of that algorithm to feedback at higher levels. Second, the statistical technique used to compute the score generally combines all of the variables into a single equation that linearly predicts the scores, often with relatively arbitrary cutoffs between scores. That is, each score is comprised of a weighted combination of all of the variables, rather than a selective subset of variables (unless a simple threshold technique with a few variables has been used). Thus, the statistical methods that are most commonly employed also render it challenging to provide meaningful (strategy-focused) feedback to the writer.

In this study, we approach this problem by asking a relatively intuitive question: how might an expert rater approach the task of giving a holistic score to an essay? When scoring an essay, with the eventual goal of providing feedback or even solely to provide a score, does the human read the essay, consider all variables simultaneously (as in a regression or massive machine learning algorithm), and then output a score? The answer to that question, based on years of research in the area of cognition, need not rely on intuition; working memory limitations, combined with the demands associated with reading and comprehension processes, problem solving, and decision making, all suggest a clear negative response. It appears unlikely, given such limitations, that all of the variables are considered simultaneously in the rater's mind in a fashion similar to a regression formula. Expert essay raters and teachers who are grading papers are likely to use a number of techniques. However, unfortunately, there is no research on this topic to our knowledge to bolster this claim. Nonetheless, based on our own experiences rating essays, we can make some educated guesses on potential approaches. A rater might begin by sorting the essays into piles. Notably, if the rater were using a regression formula (cognitively), then all of the piles would be assessed using all of the same variables. Hence, if the rater began by sorting a group of essays into "low" and "high" piles using a set of variables, then that same set of variables would be used to further sort the essays to make finer distinctions (i.e., the rater would use the same criteria for both the low and high group). This approach seems unlikely because once the

essays have been divided into clearly distinct groups, then more fine-grained categorizations ought to be based on new sets of criteria that are tailored to each group. Since expert raters are likely to have learned what those criteria are, these raters are likely to recognize certain criteria in an essay and implicitly categorize the essay as higher or lower quality while proceeding through an essay. In sum, we presume that expert raters might engage in something similar to a sorting task, initially grouping essays based on relatively superficial criteria, and subsequently classifying essays based on finer grained characteristics of the essay. These criteria are in turn likely to be related to relevant feedback that may aid the writer in revising the essay later.

We have attempted in this study to translate what we might observe in human raters within a computational algorithm by using hierarchical classification with different variables allowed to enter at each level. We can contrast this approach to a simultaneous regression of all of the variables. Our approach in this study is similar to an incremental algorithm for hierarchical classification and to hierarchical classification in general (e.g., Bianchi, Gentile, & Zaniboni, 2006; Dumais & Chen, 2000; Granitzer, 2003). However, this methodology has not been applied in the AES or AWE literature and, to our knowledge, it has not been previously used to drive feedback. In the current study, we examine the reliability of using hierarchical classification using different essay features at each stage and level of the hierarchy. In the first step, we assume that writing fluency constitutes one of the largest differences distinguishing good and poor essays. One proxy for the fluency of a writer is the length of the essay. Essays are often distinguished in terms of their length, where essays with higher scores are longer than lower rated essays (e.g., Crossley, Weston, McLain Sullivan, & McNamara, 2011; Ferris, 1994; Frase, Faletti, Ginther, & Grant, 1997; Guo, Crossley, & McNamara, 2013; Jarvis, Grant, Bikowski, & Ferris, 2003; McNamara, Crossley, & McCarthy, 2010; McNamara et al., 2013). Shorter essays have fewer words and fewer paragraphs, which in turn are indicative of less fluency on the part of the writer. We assume that longer essays will tend to have higher scores than shorter essays, but more importantly we assume the features that characterize short and long essays will be different because we presume that more fluent writers will produce more sophisticated linguistic features related to essay quality. In addition, more fluent writers most likely write essay drafts more quickly than less fluent writers. These fluent writers can then use the time remaining after completing an essay draft to focus on revising content and argumentative structure (Deane, 2013).

Hence, the first hierarchical category is determined as a function of those that meet a threshold for number of words and number of paragraphs. In the following stages of this analysis, we assume that those in the lower half (i.e., shorter essays), and those in the upper half (i.e., longer essays) will be characterized by different features, and hence, their scores will be predicted by different sets of features. Consequently, this approach requires that machine-learning algorithms be calculated separately for each group.

2. Method

2.1. Research instruments and indices

Three research instruments were used in this study, including *Coh-Metrix, the Writing Analysis Tool (WAT), and the Linguistic Inquiry and Word Count (LIWC)*. Coh-Metrix (e.g., Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Graesser, 2012; McNamara, Graesser, McCarthy, & Cai, 2014) measures text difficulty, text structure, and cohesion through the integration of lexicons, pattern classifiers, part-of-speech taggers, syntactic parsers, shallow semantic interpreters, and other components that have been developed in the field of computational linguistics (Jurafsky & Martin, 2008). Coh-Metrix reports on hundreds of linguistic variables that are primarily related to text difficulty (McNamara et al., 2014). Coh-Metrix also provides a replication of features reported by Biber (1988) including tense and aspect markers, place and time adverbials, pronouns and pro-verbs, questions, nominal forms, passives, stative forms, subordination features, prepositional phrases, adjectives and adverbs, modals, specialized verb classes, reduced forms and dispreferred structures, and coordinations and negations.

We also used the *WAT*, which we are currently developing specifically for the purpose of examining essays. The *WAT* includes a variety of variables designed specifically to assess the quality of a written document, or writing proficiency. These variables relate to global cohesion, contextual cohesion, lexical sophistication, n-gram frequency, key word use, and rhetorical features (McNamara et al., 2013).

LIWC is an automated word analysis tool developed by Pennebaker and his colleagues that reports the percentage of words in a text that are in particular psychological categories (Pennebaker, Booth, & Francis, 2007). The categories include linguistic processes (e.g. pronouns, past tense), psychological processes (e.g., social processes, cognitive processes, perceptual processes), personal constructs (e.g., work, religion), and paralinguistic dimensions (e.g., speech disfluencies). LIWC counts the number of words that belong to each word category and provides a proportion score that divides the number of words in the category by the total number of words in the text.

2.2. Cohesion

Cohesion emerges from the presence or absence of cohesive cues that tie different parts of the text together. For example, connectives provide information about the relationship between clauses, sentences, and ideas in a text. Lexical and semantic overlap between sentences provides cues that ideas are related to each other. This study included the following cohesion indices.

2.2.1.1. Connectives

Connectives increase text cohesion by explicitly linking ideas and clauses together (Crismore, Markkanen, & Steffensen, 1993; Longo, 1994). Coh-Metrix calculates the incidence score for connectives in a text as the number of occurrences per 1000 words. In addition to calculating a measure of all connectives contained within a given text. Coh-Metrix contrasts positive (*also, moreover*) versus negative (*however, but*) connectives and provides indices on five categories of connectives (Halliday & Hasan, 1976; Louwerse, 2001): causal (*because, so)*, contrastive (*although, whereas*), additive (*moreover, and*), logical (*or, and*), and temporal (*first, until*). Based on Biber (1988), Coh-Metrix also provides the incidence score for the connective *because* (i.e., an *unambiguous* causative adverbial subordinator) and *other subordinators* (e.g., although, while, how), which can serve to connect subordinate clauses.

2.2.1.2. Lexical and semantic co-referentiality

Research has shown that lexical and semantic overlap aid in text comprehension (Douglas, 1981; Kintsch & van Djik, 1978; Rashotte & Torgesen, 1985). Coh-Metrix considers four forms of lexical overlap between sentences: noun overlap, argument overlap, stem overlap, and content word overlap (McNamara, Crossley, et al., 2010; McNamara, Louwerse, et al., 2010; McNamara et al., 2014). Latent Semantic Analysis (LSA) is used to measure the semantic co-referentiality between sentences and paragraphs in a text (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). LSA is a statistical representation of deeper world knowledge used to assess the level of semantic cohesion within a text. LSA utilizes singular value decomposition to condense large corpora of texts into approximately 300–500 dimensions. These dimensions reflect the relative frequency of words' occurrence within given texts or larger sections of text and represent the degree of semantic similarity between words – an important indicator of text cohesion (Landauer, McNamara, Dennis, & Kintsch, 2007). In addition to these separate indices. Coh-Metrix also provides a referential cohesion component score that combines lexical and semantic overlap indices based on a principal component analysis (e.g., Graesser, McNamara, & Kulikowich, 2011; McNamara et al., 2014).

2.2.1.3. Causal cohesion

Coh-Metrix calculates the level of causal cohesion in a text by measuring the ratio of causal verbs to causal particles (Graesser et al., 2004; Dufty, Hempelmann, Graesser, Cai, & McNamara, 2005). The measure of causal verbs is based on the frequency count of main clausal verbs identified through WordNet (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). The causal particles are counted based on a defined set of main causal verbs, such as *because* and *as a result*. Causal cohesion can reduce text comprehension difficulties as it reveals causal relationships between simple clauses, as well as between events and actions (Pearson, 1974–1975).

2.2.1.4. Lexical diversity

Traditional lexical diversity indices typically measure the ratio of types (i.e., unique words occurring in the text) by tokens (i.e., all instances of words) where higher numbers (from 0 to 1) indicate greater lexical diversity (Templin, 1957). Lexical diversity is at a maximum when all of the words in a text are different, or the number of word types is equal to the total number of words (tokens). In that case, the text is likely to be either very low in cohesion or very short. By contrast, lexical diversity is lower (and cohesion is higher) when words tend to be used multiple times across a text. Coh-Metrix provides several lexical diversity indices, including type-token ratio and D (Malvern, Richards, Chipere, & Durán, 2004).

2.2.1.5. Spatiality

Spatial information aids in the construction of a well-structured situational model (Kintsch & van Djik, 1978) that clearly conveys the meaning of the text. Spatiality and spatial cohesion is measured using two forms of information: location information and motion information (Dufty, Graesser, Lightman, Crossley, & McNamara, 2006; Dufty, Graesser, Louwerse, & McNamara, 2006). Both motion verbs and location nouns are identified through WordNet (Fellbaum, 1998; Miller et al., 1990). Spatial and place adverbials provide information about position and direction in space (e.g., inside, away, north, indoors).

2.2.1.6. Temporality

Temporality and temporal cohesion is measured by Coh-Metrix in five ways: temporal adverbials, aspect repetition, tense repetition, the combination of aspect and tense repetition, and a temporal cohesion component score. Temporal adverbials provide information about duration and position in time (e.g., once, afterwards, simultaneously, yesterday). Time is represented in text through two dimensions: tense (past, present, future) and aspect (in progress versus completed). With the use of these dimensions, Coh-Metrix calculates the consistency of tense and aspect across a passage of text. As shifts in tense and aspect occur, the Coh-Metrix repetition score decreases. Thus, a low score indicates that the representation of time in a given text may be disjointed, which could have a negative consequence on the construction of a mental representation. The temporal cohesion component score combines tense and aspect repetition indices based on a principal component analysis (e.g., Graesser et al., 2011; McNamara et al., 2014).

2.2.1.7. Paragraph cohesion

WAT calculates the lexical and semantic overlap between paragraphs (initial to middle paragraphs, middle paragraphs to final paragraph, and initial paragraph to final paragraph) and between the essay prompt and the essay. The semantic similarity among the paragraphs, essay, and prompt are calculated using LSA cosine values. Lexical overlap, on the other hand, is calculated using measures of key word overlap. High lexical and semantic overlap between paragraph types is related to judgments of essay coherence and quality (Crossley & McNamara, 2011; Crossley, Roscoe, McNamara, & Graesser, 2011).

2.3. Vocabulary

The vocabulary, or the particular words used within an essay, is strongly related to the perceived quality of the essay (McNamara, Crossley, et al., 2010). Higher quality essays are associated with the use of less frequent words and phrases that are less familiar to most readers, more concrete and imageable words, less ambiguous words, and words with more specific meanings (Crossley, Weston, et al., 2011; McNamara et al., 2013).

2.3.1.1. Word frequency

Coh-Metrix calculates word frequency using the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995), which consists of word frequencies computed from a 17.9 million-word corpus. Word frequency indices measure how often particular words occur in the English language. Word frequency is an important indicator of lexical knowledge. Coh-Metrix calculates word familiarity using the MRC Psycholinguistic Database (Coltheart, 1981). Word familiarity refers to words that are more readily

recognized, but not necessarily more frequent (e.g., compare *eat* to *while*). *WAT* also provides frequency counts for words in the Academic Word List (AWL: Coxhead, 2000). The AWL consists of 3110 words commonly found in academic writing (including 570 headwords or lemmas). The use of more academic words is associated with less frequent and more specific words.

2.3.1.2. Flesch Kincaid grade level

Flesch Kincaid grade level (Kincaid, Fishburne, Rogers, & Chissom, 1975) provides a metric of text difficulty based on the length of words and sentences within the text. In Coh-Metrix, the Flesch Kincaid grade level (RDFKGL) is computed as [(.39 × sentence length) + (11.8 × word length) − 15.59] where word length is measured as the mean number of syllables per word. Because this index is based primarily on the type and number of words used in an essay, it is generally correlated with the author's vocabulary knowledge.

2.3.1.3. Word information measures

Word information indices are calculated in Coh-Metrix using the MRC Psycholinguistic Database and WordNet (Fellbaum, 1998; Miller et al., 1990; Wilson, 1988), including measures of concreteness, imageability, meaningfulness, hypernymy, and polysemy. Higher concreteness scores typically refer to words that reference objects, materials, or persons. Imageability refers to the ease with which words evoke mental images. Word meaningfulness indices measure how strongly words associate with other words. A hypernymy value represents the degree of specificity of a word within a conceptual hierarchy. Thus, a low hypernymy score reflects a text that is more vague. Polysemy refers to the number of senses a word has; ambiguous words have more senses. Thus, word polysemy is indicative of the level of text ambiguity. These measures are associated with word knowledge and lexical sophistication (Crossley, Salsbury, & McNamara, 2010).

2.3.1.4. N-grams

WAT compares the normalized frequency of n-grams (i.e., sequences of words such as bigrams and trigrams) shared in both a reference corpus taken from the written and spoken British National Corpus (BNC) and the language sample of interest (e.g., the essays used in this study). The indices report the frequency of the n-grams within the essay, correlations that represent the similarity between the frequency of occurrences in the reference corpus and the essay, and the proportion of n-grams (relative to the total number of words) within the essay. Crossley, Cai, and McNamara (2012) found that higher proficiency writers use less frequent (i.e., rarer) n-grams and show greater similarity (i.e., higher correlations) in terms of n-gram frequency as compared to the representative corpus. More proficient writers also produce essays that contain proportionally fewer n-grams. Hence, better essays include less commonly used language.

2.3.2. Nominals

We measure a number of different linguistic features related to noun phrases in order to assess noun density, referentiality, specificity, and modification. These include incidence scores for different types of noun phrase heads such as nouns (e.g., singular and plural nouns) and pronoun types (e.g., all pronouns, 2nd person pronouns, 3rd person pronouns). We also measure incidence scores for non-head elements of noun phrases such as determiners, demonstratives, and adjectives. In addition, we measure linguistic features for embedded clauses in noun phrases (e.g., relative clauses). Finally, we measure the incidence of noun phrases to include sentence relative clauses that function as noun phrases (e.g., What I really want is to eat). Features of nouns phrases are an important indicator of writing quality with higher level essays containing more modifiers per noun phrase (Crossley, Weston, et al., 2011), higher rated essays containing more nominalizations (Guo et al., 2013), and lower rated essays containing more personal pronouns (Crossley, Roscoe, et al., 2011).

2.3.3. Verb-related features

We calculated a variety of indices related to a text's verb content to investigate verb density, verbal semantics, and verb forms. These indices include features such as incidence of verbs, incidence of verb phrases, incidence of auxiliary verbs, incidence of participles (present and past), and incidence of verb

base forms. We also calculated verbal semantic indices such as incidence of private verbs (e.g., doubt, know, fear), incidence of public verbs (e.g., say, deny, examine), and incidence of 'be' verbs. Verbs are an important indicator of writing quality with lower rated essays containing more verb base forms (Crossley, Roscoe, et al., 2011) and more verbs in present tense. In contrast, more highly rated essays contain past participle verbs (Guo et al., 2013).

2.3.4. Syntactic indices

Coh-Metrix calculates a number of indices that measure syntactic complexity. These indices include the mean number of words before the main verb, the mean number of higher-level constituents (sentences and embedded sentence constituents) per word, the average number of modifiers per noun phrase, incidence of embedded clauses, incidence of 'that' deletion, and the incidence of infinitives. Coh-Metrix also provides a syntactic complexity component score that combines syntactic indices based on a principal component analysis (e.g., Graesser et al., 2011; McNamara et al., 2014). Additional measures related to syntax include frequency of punctuation such as commas, periods, and semicolons. Coh-Metrix also measures syntactic similarity by calculating, for a given text, the uniformity and consistency of the syntactic constructions at the clause, phrase, and word level. Higher rated essays contain more complex syntactic structures, which is measured by indices such as the number of words before the main verb (McNamara, Crossley, et al., 2010) and the number of embedded clauses (Guo et al., 2013).

2.3.5. Rhetorical and semantic features

2.3.5.1. Paragraph specific n-grams. WAT reports on a variety of n-gram indices that are specific to the uses of word sequences within sections of the essays (i.e., introduction, body, and conclusion paragraphs). The use of higher quality paragraph n-grams has shown positive correlations with writing proficiency (Crossley, Roscoe, et al., 2011; McNamara et al., 2013).

2.3.5.2. Lexical features. WAT reports on a variety of lexical categories related to rhetorical style. These include amplifiers (e.g., completely, extremely), hedges (e.g., almost, maybe), indirect pronouns, assent/agreement terms (e.g., agree, absolutely, yes), exemplification (e.g., for instance, for example), prediction modals (will, shall, wouldn't), downtoners (e.g., barely, somewhat), negations (not, neither, nor), and the use of the terms seem/appear.

2.3.5.3. Psychological semantics. LIWC reports on a variety of psychological word categories including cognitive processes (e.g., think, know), perceptual processes (e.g., hear, feel), social processes (e.g., talk, mate), and religious terms (e.g., alter, mosque). These indices can provide information about the psychological states of writers or speakers and may be predictive of human judgments of language proficiency.

2.3.5.4. Narrativity. The degree of narrativity versus informational content provided within the essay is assessed using the narrativity component score provided by Coh-Metrix (Graesser et al., 2011; McNamara, 2013). Essays with high narrativity are more story-like, with events, places, and things that are more common in everyday life. Narrative is closely affiliated with everyday, oral conversation. Essays with lower narrativity include more content (e.g., nouns) and discuss less familiar topics. Using more information as evidence within an essay is associated with more refined rhetorical strategies on the part of the writer, and thus higher quality essays tend to have lower narrativity scores.

2.4. Corpus selection

Our corpus consisted of 1243 argumentative (persuasive) essays. Because our interest is in developing a general algorithm that is predictive across a broad range of prompts, grade levels, and temporal conditions, we selected a general corpus that contained 14 different prompts, 3 different grade levels (9th grade, 11th grade, and college freshman), and four different time limits for writing or *temporal* conditions (essays that were untimed and essays that were written in 10, 15, and 25 minute increments). The essays also came from a variety of geographical locations: (Tennessee, Mississippi, Florida, the District of Columbia, and New York).

Table 2
Corpus description.

Corpus name	Prompts	Grade levels	Published in	Timing	*n*
High School 1	2	9th and 11th	Crossley, Weston, et al. (2011)	25 minutes	101
High School 2	1	9th and 11th	Weston, Roscoe, Floyd, and McNamara (2013)	15 minutes	353
High School 3	1	11th	Crossley, Weston, et al. (2011)	25 minutes	70
College Board	1	12th	Obtained from College Board	25 minutes	40
Mississippi 1	3	Freshman college	McNamara, Crossley, et al. (2010); Crossley and McNamara (2010)	Untimed	184
Mississippi 2	2	Freshman college	Crossley, Roscoe, et al. (2011); Crossley and McNamara (2011)	25 minutes	315
University of Miami	1	Freshman college	Wolfe et al. (2009)	Unknown	59
Memphis 1	2	Freshman college	Crossley, White, McCarthy, and McNamara (2009)	25 minutes	70
Memphis 2	1	Freshman college	Raine et al. (2011)	10 minutes	51

Notably, for this study, we selected a wide distribution of prompts, grade levels, temporal conditions, and geographical locations. This approach contrasts with one in which an algorithm is developed for a target prompt, grade level, or essay writing conditions (such as time restrictions). Developing such condition-specific algorithms is necessary in certain contexts, particularly when prompts call for vastly different writing styles (see Ramineni & Williamson, 2013, for an overview of prompt-specific and generic scoring models). However, our question regarded the more general question of the feasibility of a hierarchical approach to developing an automated algorithm. Hence, our goal was to develop an algorithm with the potential of generalizing across a variety of conditions such as the specific content of the prompts, the writers' population-specific ability levels, or the particular constraints of the situation. Although we use a number of different prompts in this study, it is important to note that they are all relatively similar in their style and requirements (i.e., they are all timed, SAT-style, argumentative essay prompts). It is likely that this scoring model would not generalize to vastly different writing tasks, such as source-based writing.

The majority of the essays used in this study have been used in previous studies that have focused on writing quality (Crossley & McNamara, 2010, 2011; Crossley, Weston, et al., 2011; Crossley, Roscoe, et al., 2011; McNamara, Crossley, et al., 2010; Raine, Mintz, Crossley, Dai, & McNamara, 2011; Wolfe, Britt, & Butler, 2009). Descriptive statistics for the corpus are provided in Table 2. Differences between the corpora based on text length are provided in Table 3. As expected, the essays that were written given shorter time limits included fewer words. The inclusion of these essays was expected to provide a broader spectrum of writing quality for shorter essays (than would be observed for essays provided 25 minutes to complete) because the time limit would restrict the length of the essays but would not necessarily affect associated features of writing (e.g., writing sophistication associated with writing ability).

2.5. Human judgments

Each essay in the corpus was scored independently by two or three expert raters using a 6-point rating scale developed for the SAT. This rubric is designed to be generic and thus can be used to assess the overall quality of any argumentative essay. Importantly, this means that the rubric is not tied to

Table 3
Descriptive statistics for number of words as a function of time limit conditions.

Time limits	Mean	Median	Standard deviation
10 minutes	182.941	172.000	61.141
15 minutes	207.378	204.000	71.681
25 minutes	325.379	315.500	126.991
Unknown	347.586	332.000	165.162
Unlimited	729.745	759.500	134.991

a specific writing prompt or topic – rather, the criteria for determining quality are related to general essay properties, such as sophisticated vocabulary and evidence-based reasoning (see Appendix). The rating scale was used to holistically assess the quality of the essays and had a minimum score of 1 and a maximum score of 6. Raters were first trained to use the rubric with a small sample of argumentative essays. A Pearson correlation analysis was used to assess inter-rater reliability between raters. When the raters reached a correlation of $r = .70$, the ratings were considered reliable and the raters scored a larger subsection of the corpus.

The raters reported exact matches between essay scores for 627 out of the 1243 of the essays (50% exact accuracy; $df = 25$, $n = 1243$, $\chi^2 = 1364.978$, $p < .001$, $r = .750$, $p < .001$, Kappa = .519). The raters reported adjacent matches for 1173 of the 1243 essays (94% adjacent accuracy). We used the mean score rounded down from the raters as the holistic value for the quality of each essay (i.e., an essay that had an average score 2.5 was given a final score of 2). We also used the mean score rounded down because only 8 of the 1243 essays had a score of 5.5 or higher. Since this score category is rare, developing a model to predict it is not parsimonious.

3. Results

3.1. Statistical analysis

To select the variables for use in this analysis, we first conducted correlations between the variables reported by Coh-Metrix, *WAT*, and LIWC ($N = 440$) and the human scores for each essay. We selected each variable that demonstrated a significant correlation ($p < .050$) with essay scores and also showed, at minimum, a weak effect size ($r \geq .10$) for differences as a function of essay score. This yielded 320 variables. These indices were then checked for multicollinearity (r between any two variables >.70).[1] If two or more variables demonstrated multicollinearity, then the variable with the highest correlation to essay score was retained and the other variable(s) was removed from the analysis. This procedure resulted in removing 180 variables from the analysis leaving us with 140 variables for the analysis.

Next, for each partition of the data, we conducted a stepwise discriminant function analysis (DFA) using the 140 indices (i.e., the indices that significantly correlated with essay score but did not demonstrate multicollinearity). The DFA generates a discriminant function, which acts as an algorithm to predict group membership (i.e., the proficiency level of the writers). We selected a DFA based on the results of Jarvis (2011) who found that DFAs were superior to other machine learning techniques (e.g., Support Vector Machines, Naïve Bayes classifiers, Cart classifiers) in a similar corpus classification task. We use the DFA first on the entire partition of essays and then the DFA model reported from the entire set is used to predict group membership of the essays using leave-one-out-cross-validation (LOOCV). In LOOCV, a fixed number of folds equal to the number of observations (i.e., texts) is selected. Each fold is then used to test the model such that one observation in turn is left out and the remaining instances are used as the training set (in this case the remaining essays). The accuracy of the model is tested on the model's ability to predict the proficiency classification of the omitted instance. This allows us to test the accuracy of the model on an independent data set. If the results of the discriminant analysis in both the entire set and the *n*-fold cross-validation set are similar, then the findings support the extension of the analysis to external data sets.

We report the findings of the DFA using an estimation of the accuracy of the analysis. This estimation is made by plotting the correspondence between the classifications of the essays (either low or high scored essays for each partition) and the predictions made by the DFA model. For each partition, we also assess fives types of accuracy with the human scores: chi-square, Pearson *r*, Cohen's Kappa, exact accuracy, and adjacent accuracy. Exact accuracy examines how accurate the DFA model is at assigning the same score to the essay as did the human raters. Adjacent accuracy examines how accurate the DFA model is at assigning a score to the essay that is either exactly the same or adjacent to that assigned by the human raters. Thus, if the model assigned a score of 4 to an essay that was

[1] Multicollinearity between indices indicates that the indices are effectively measuring the same patterns in the data.

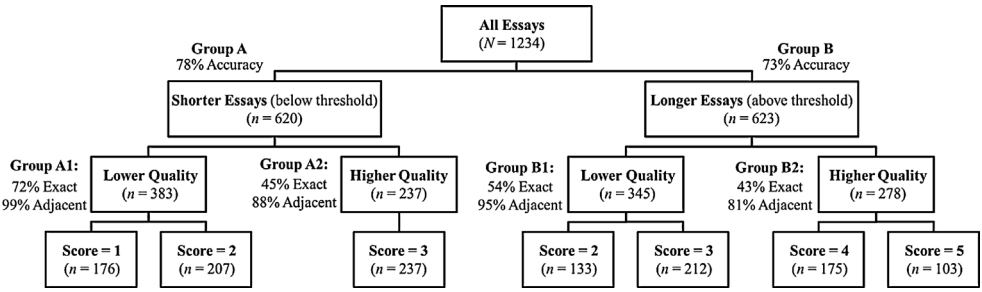

Fig. 1. Graphical depiction of hierarchical approach to predicting essay scores. First essays are divided into two sets based on a length threshold. Then algorithms are used to further divide each subset into lower and higher quality essays. The specific scores are then predicted based on algorithms within the third level of the hierarchy.

scored by the human raters as a 4, the exact accuracy would 1 and the adjacent accuracy would 1. If the model assigned the same essay a score of 3 (or 5), the exact accuracy would be 0 and the adjacent accuracy would be 1. If the model assigned the essay a score of 2 (or 6), both the exact accuracy and the adjacent accuracy would be 0.

The unique approach that we adopt here is to use hierarchical classification rather than predicting the score with one algorithm or model. Thus, the essay scores are iteratively predicted. First, all of the essays are divided into two sets (see Fig. 1), those that fall below a length criteria (i.e., shorter essays, Group A) and those that meet or exceed a length criteria (i.e., longer essays, Group B). The assumption is that these two initial sets of essays should be judged on different criteria. Thus, separate analyses are iteratively conducted until each of the scores is predicted. In the case of this data set, there were only 3 essays with a score of 6 (and 5 essays with a score of 5.5). Thus, we predicted 5 essay scores (i.e., 1–5).

3.2. Shorter essays (Group A)

Those essays with 250 words or fewer or only 1 or 2 paragraphs were categorized as *shorter essays* that did not meet a minimum length threshold. We selected the length threshold based on the average word-lengths for the essays in this corpus composed within 10, 15, and 25-minute time limits ($n = 1062$ essays; $M = 273.802$, $SD = 122.285$), rounding the average score to the nearest fiftieth (i.e., 250 words). Our starting assumption is that the quality of the shorter essays will be predicted by a *different* algorithm than that used to predict the quality of longer essays with more words and paragraphs. We imposed the paragraph threshold following the assumption that persuasive essays include an introduction, body, and conclusion, which require a minimum of three paragraphs. This set included 620 essays (49.9%) with human scores ranging from 1 to 4. None of the essays that fell below these minimum thresholds were assigned scores of 5 or 6 by the expert raters.

3.2.1. Shorter essays: lower and higher relative quality partitions

3.2.1.1. Discriminant function analysis. Shorter essays were subsequently divided into lower relative quality essays (Group A1) and higher quality essays (Group A2) based on a DFA. The stepwise DFA selects variables based on a statistical criterion that retains the variables that best classify the grouping. For our analysis, the significance level for a variable to enter or to be removed from the model was set at the $p \leq .05$. The stepwise DFA retained 14 variables as significant predictors of either lower or higher quality essays (see Table 4 for descriptive statistics and f and p values) and removed the remaining 126 variables as non-significant predictors.

The results demonstrate that the DFA using the 14 significant indices correctly allocated 484 of the 620 essays in the total set ($df = 1$, $n = 620$, $\chi^2 = 172.498$, $p < .001$, $r = .527$, $p < .001$, Kappa = .502) for an accuracy of 78.1% (chance for this analysis and all analyses is 50%). In the LOOCV, the DFA correctly allocated 476 of the 620 essays for an accuracy of 76.8%. The confusion matrix provided in Table 5

Table 4
Descriptive statistics for indices predicting Group A1 and A2 essays.

Index	Index type	Group A1 (essays scored 1–2)	Group A2 (essays scored 3–4)	*f*
Nominalizations	Nominals	3.308 (3.355)	6.058 (4.520)	64.510
Plural nouns (incidence)	Nominals	75.458 (33.788)	94.436 (35.491)	39.399
Lexical diversity (*D* score)	Cohesion	73.706 (32.780)	92.228 (23.418)	34.874
Present participle (incidence)	Verb-related	18.620 (13.903)	14.140 (9.834)	30.937
Social processes	Rhetorical/semantic	12.334 (4.291)	10.214 (3.240)	28.726
Third person pronouns	Nominals	6.445 (5.600)	8.902 (7.010)	26.541
LSA paragraph-to-paragraph (mean)	Cohesion	0.156 (0.212)	0.249 (0.234)	24.988
Religious terms	Rhetorical/semantic	0.126 (0.363)	0.119 (0.319)	23.498
Synthetic negation	Rhetorical/semantic	0.032 (0.176)	0.013 (0.113)	22.294
Commas (incidence)	Syntactic	29.813 (22.552)	40.673 (20.393)	21.218
Sentence relative clauses	Nominals	0.188 (0.581)	0.137 (0.414)	20.175
Flesch Kincaid grade level	Vocabulary	8.240 (2.2148)	9.262 (2.006)	19.154
Place adverbials	Cohesion	0.498 (0.906)	0.908 (1.387)	19.426
Assent/agreement terms	Rhetorical/semantic	0.197 (0.463)	0.080 (0.202)	18.419

Note: All $p < .001$.

indicates that for the total set of essays, 26 (17%) of the essays scored more highly (i.e., given a 3 or 4) by the expert raters were misclassified as lower quality essays, and 110 (24%) of the lower quality essays (i.e., given a 1 or 2 by the expert raters) were (mis)estimated to be of higher quality by the algorithm.

3.2.2. *Classifying Group A1 essays*

Shorter essays that were partitioned by the initial DFA as lower quality (Group A1) were then further partitioned into essays predicted to have been scored as 1 versus those essays predicted to have been scored as 2 by the expert raters. Of these, the raters assigned 181 of the essays a score of 1, 176 of the essays a score of 2, 25 of the essays a score of 3, and 1 essay a score of 4. Hence, the best that this model can do (i.e., at this level of the hierarchy) is to assign a 1 to those that received a 1 by the human raters, and a 2 to those that received a 2, 3, or 4.

3.2.2.1. Discriminant function analysis. A stepwise DFA was conducted to classify the Group A1 (see Fig. 1) essays scored as either 1 or 2. The stepwise DFA retained 15 variables as significant predictors of essay score (see Table 6 for descriptive statistics, *f* and *p* values) and removed the remaining 125 variables as non-significant predictors.

The results demonstrate that the DFA using the 15 significant indices correctly allocated 300 of the 383 essays in the total set ($df = 1$, $n = 383$, $\chi^2 = 122.195$, $p < .001$, $r = .565$, $p < .001$, Kappa = .565) for an accuracy of 78.3%. In the LOOCV, the DFA correctly allocated 291 of the 383 essays for an accuracy of 76.0%.

The confusion matrix provided in Table 7 shows that for the total set of essays, 44 of the essays scored as a 1 by humans were assigned a score of 2 by the algorithm, and 39 essays given a score of 2 or 3 by the human raters were assigned a lower score (i.e., 1) by the algorithm. In sum, 276 of the 383

Table 5
Confusion matrix for Group A (i.e., shorter) essays showing actual and predicted essay quality.

Actual text type	Predicted text type	
	Lower quality essays	Higher quality essays
Lower quality essays	357	110
Higher quality essays	26	127

Table 6
Descriptive statistics for indices predicting Group A1 (i.e., shorter, lower quality) essays as a function of the low and high partitions.

Index	Index type	Low partition (essays scored 1)	High partition (essays scored 2)	*f*
LSA paragraph-to-paragraph (standard deviation)	Cohesion	0.008 (0.032)	0.032 (0.066)	34.247
Other subordinators	Cohesion	0.442 (1.165)	0.203 (0.492)	29.017
Seem/appear	Rhetorical/semantic	0.072 (0.333)	0.312 (0.703)	26.109
Demonstratives	Nominals	0.834 (1.078)	1.406 (1.540)	24.052
Narrativity score (*Z* score)	Rhetorical/semantic	1.085 (0.709)	0.747 (0.582)	22.721
That deletion	Syntactic	0.591 (0.829)	0.837 (1.096)	21.208
Type-token ratio	Cohesion	58.930 (9.010)	54.693 (7.532)	20.056
Perceptual processes	Rhetorical/semantic	2.539 (2.444)	3.026 (2.1667)	19.516
Lexical diversity (*D* score)	Cohesion	59.320 (38.069)	79.445 (23.0567)	18.354
Prediction modals	Rhetorical/semantic	1.834 (2.287)	1.668 (1.717)	19.092
Public verbs	Verb-related	8.109 (10.344)	4.742 (5.969)	17.946
Because (incidence)	Cohesion	0.746 (1.033)	0.718 (1.005)	16.992
Infinitives	Syntactic	4.232 (3.311)	5.535 (3.983)	16.200
Conclusion n-grams	Rhetorical/semantic	0.011 (0.104)	0.045 (0.207)	15.504
Cognitive processes	Rhetorical/semantic	23.0645 (5.695)	22.473 (4.244)	14.919

Note: All $p < .001$.

essays reported exact matches in scores (72% exact accuracy) and 299 of the essays reported adjacent matches in scores (99% adjacent accuracy).

3.2.3. Classifying Group A2 essays

Further analyses of shorter essays focused on those partitioned as higher quality (Group A2, Fig. 1) were further partitioned into essays predicted to have been scored 3 and essays predicted to have been scored 4. Of these, the human raters assigned 28 of the essays a score of 1, 82 a score of 2, 107 a score of 3, and 20 of the essays a score of 4.

The outcomes from the DFA analysis conducted for these essays when compared to the human ratings for the 237 essays resulted in 107 exact matches (45% exact accuracy) and 192 adjacent matches (81% adjacent accuracy). We do not report the specifics of this DFA because higher accuracies were possible for this partition of essays by simply assigning each essay a score of 3. That is, given the structure for the range of essay scores (i.e., 1–5), the best possible performance at this level of the hierarchy was to not separate the essays into higher and lower quality scores. In this case, 107 of the 237 essays reported exact accuracy (45% exact accuracy) and 209 of the 237 essays reported adjacent accuracy (88%). Hence, this was the model that we used.

3.2.4. Overall accuracy of scoring models for shorter essays

In summary, the models developed for the essays that did not meet the length threshold reported exact matches between the predicted essay scores and the human scores for 383 out of the 620 of the essays (62% exact accuracy; $df = 6$, $n = 620$, $\chi^2 = 334.068$, $p < .001$, $r = .620$, $p < .001$, Kappa = .521). The models reported adjacent matches for 589 of the 620 essays (95% adjacent accuracy).

Table 7
Confusion matrix for the total set of Group A1 (i.e., shorter, lower quality) essays showing actual and predicted essay scores.

Actual essay score	Predicted essay score	
	1	2
1	137	44
2	37	139
3	2	23
4	0	1

Table 8
Descriptive statistics for indices predicting Group B1 and B2 essays.

Index	Index type	Group B1 (essays scored 1–3)	Group B2 (essays scored 4–5)	*f*
Bigram frequency (spoken)	Vocabulary	0.006 (0.002)	0.005 (0.002)	81.714
Commas (incidence)	Syntactic	35.884 (18.627)	45.476 (17.704)	54.950
Public verbs	Verb-related	1.328 (1.668)	1.222 (1.522)	43.116
Temporal adverbials	Cohesion	2.854 (3.276)	3.100 (3.088)	35.611
Nominalizations	Nominals	10.346 (7.439)	16.536 (11.295)	31.116
Adjectives (incidence)	Nominals	69.723 (21.294)	77.851 (21.043)	27.960
Auxiliary verbs (incidence)	Verb-related	2.959 (4.214)	2.062 (2.609)	25.077
Verb base form (incidence)	Verb-related	42.535 (16.159)	35.467 (13.167)	22.964
Amplifiers	Rhetorical/semantic	0.997 (1.329)	1.418 (1.678)	21.327
WH relative clauses in subject position	Nominals	1.190 (1.746)	1.782 (2.285)	19.797
Pronoun you (incidence)	Nominals	0.765 (1.512)	0.356 (0.990)	18.541
Other subordinators	Cohesion	1.297 (2.447)	0.987 (1.736)	17.618
Private verbs	Verb-related	12.515 (8.274)	9.406 (6.0431)	16.700
Sentence relative clauses	Nominals	0.406 (0.783)	0.444 (0.867)	15.938

Note: All $p < .001$.

3.3. Longer essays (Group B)

Essays with 251 or more words and with 3 or more paragraphs were categorized as longer essays (Group B, see Fig. 1). This set included 623 essays (51.1%) with human scores ranging from 1 to 6; however, only 3 essays were scored as a 6 by the human raters in this data set.

3.3.1. Longer essays: lower and higher relative quality partitions

3.3.1.1. Discriminant function analysis. As in the earlier analysis of shorter essays, longer essays were partitioned into lower relative quality essays (Group B1, see Fig. 1) and higher relative quality essays (Group B2) based on a DFA. The stepwise DFA retained 14 variables as significant predictors of either lower or higher quality essays (see Table 8 for descriptive statistics and *f* and *p* values) and removed the remaining 126 variables as non-significant predictors.

The results demonstrate that the DFA using the 14 significant indices correctly allocated 454 of the 623 essays in the total set ($df = 1$, $n = 623$; $\chi^2 = 124.614$, $p < .001$, $r = .447$, $p < .001$, Kappa = .444; see Table 9 for the confusion matrix) for an accuracy of 72.9%. In the LOOCV, the DFA correctly allocated 447 of the 623 essays for an accuracy of 71.7%. The confusion matrix provided in Table 9 indicates that for the total set of essays, 65 (27%) of the essays scored more highly (i.e., given a 4 or 5) by the expert raters were misclassified as *lower quality essays*, and 104 (27%) of the lower quality essays (i.e., given a 2 or 3 by the expert raters) were (mis)estimated to be of higher quality by the algorithm.

3.3.2. Classifying Group B1 essays

Longer essays that were partitioned by the initial DFA as lower relative quality essays (Group B1, see Fig. 1) were further partitioned into essays scored 2 and essays scored 3. Of these, the human raters assigned 14 of the essays a score of 1, 102 of the essays a score of 2, 164 of the essays a score of 3, 60 of the essays a score of 4, and 5 of the essays a score of 5. Hence, the best performance that can be

Table 9
Confusion matrix for the total set of Group B (i.e., longer essays) showing actual and predicted essay quality.

Actual text type	Predicted text type	
	Lower quality essays	Higher quality essays
Lower quality essays	280	104
Higher quality essays	65	174

Table 10
Descriptive statistics for indices predicting Group B1 (i.e., longer, lower quality) essays a function of the low and high partitions.

Index	Index type	Low partition (essays scored 2)	High partition (essays scored 3)	*f*
Pronoun you (incidence)	Nominals	1.410 (1.967)	0.594 (1.345)	20.523
Academic words	Vocabulary	22.431 (14.570)	28.160 (13.718)	14.142
Split infinitives	Syntactic	0.043 (0.204)	0.131 (0.351)	11.485
Aspect repetition	Cohesion	0.866 (0.125)	0.882 (0.120)	10.079
Logical connectives (incidence)	Cohesion	55.704 (17.369)	49.920 (14.149)	9.397
Verb base form (incidence)	Verb-related	49.294 (16.971)	42.154 (15.413)	8.862
Trigram frequency (written)	Vocabulary	0.066 (0.032)	0.059 (0.028)	8.352
That deletion	Syntactic	1.405 (1.626)	1.070 (1.537)	8.048
Temporal cohesion	Cohesion	5.655 (1.620)	5.443 (1.562)	7.708

Note: All $p < .001$.

expected would be for the algorithm to assign a 2 to those essays with scores of 1 or 2, and a 3 to those essays scored as a 3, 4, or 5.

3.3.2.1. Discriminant function analysis. A stepwise DFA was conducted to classify Group B1 essays scored either 2 or 3. The stepwise DFA retained 9 variables as significant predictors of essay score (see Table 10 for descriptive statistics and *f* and *p* values) and removed the remaining 131 variables as non-significant predictors.

The results demonstrate that the DFA using the 9 significant indices correctly allocated 250 of the 345 essays in the total set ($df = 1$, $n = 345$; $\chi^2 = 57.131$, $p < .001$, $r = .407$, $p < .001$, Kappa = .405) for an accuracy of 72.5%. In the LOOCV, the DFA correctly allocated 243 of the 345 essays for an accuracy of 70.5%.

The confusion matrix provided in Table 11 shows that for the total set of essays, among the 14 essays in this set scored as a 1 by the human raters, 9 were assigned a score of 2 and 5 were assigned a score of 3 by the algorithm. Of those given a score of 2, 68 were correctly classified and 34 were classified as a 3. Of those scored given a score of 3, 118 were correctly classified and 46 were classified as a 2. Of those essays given a score of 4 or 5 by human raters, 55 were scored as a 3, and 10 were scored as a 2 (see Table 11). Hence, 186 of the 345 essays reported exact matches in scores (54% exact accuracy) and 326 of the essays reported adjacent matches in scores (95% adjacent accuracy).

3.3.3. Classifying Group B2 essays

Longer essays that were partitioned by the initial DFA as higher relative quality essays (Group B2, see Fig. 1) were further partitioned into essays scored 4 and essays scored 5. Of these, the human raters assigned 3 of the essays a score of 1, 30 of the essays a score of 2, 71 of the essays a score of 3, 115 of the essays a score of 4, 56 of the essays a score of 5, and 3 of the essays a score 6. Hence, the best performance that can be expected would be for the algorithm to assign a 4 to those essays with scores of 1, 2, 3, or 4, and a 5 to those essays scored as a 5 or 6.

3.3.3.1. Discriminant function analysis. A stepwise DFA using the same criteria as used previously was conducted to classify Group B2 essays scored either 4 or 5. The stepwise DA retained 8 variables as

Table 11
Confusion matrix for the total set of Group B1 (i.e., longer, lower quality) essays showing actual and predicted essay scores.

Actual essay score	Predicted essay score	
	2	3
1	9	5
2	68	34
3	46	118
4	9	51
5	1	4

Table 12
Descriptive statistics for indices predicting Group B2 (i.e., longer, higher quality) essays as a function of the low and high partitions.

Index	Index type	Low partition (essays scored 4)	High partition (essays scored 5)	*f*
Nominalizations	Nominals	16.100 (9.906)	23.373 (13.381)	21.349
LSA paragraph-to-paragraph (standard deviation)	Cohesion	0.106 (0.053)	0.127 (0.063)	14.620
Past participles	Verb-related	24.140 (10.272)	27.866 (10.653)	11.651
Introduction n-grams	Rhetorical/semantic	0.448 (0.767)	0.746 (0.882)	9.483
Amplifiers	Rhetorical/semantic	1.406 (1.685)	1.814 (1.756)	8.736
CELEX frequency logarithm (mean for content words)	Vocabulary	2.402 (0.143)	2.341 (0.135)	8.447
WH relative clauses in subject position	Nominals	1.708 (2.096)	2.492 (3.109)	8.258
Bigrams correlation (written)	Vocabulary	0.281 (0.191)	0.268 (0.1721)	7.871

Note: All $p < .001$.

significant predictors of essay score (see Table 12 for descriptive statistics and *f* and *p* values) and removed the remaining 132 variables as non-significant predictors.

The results demonstrate that the DFA using the 8 significant indices correctly allocated 200 of the 278 essays in the total set ($df = 1$, $n = 278$; $\chi^2 = 37.419$, $p < .001$, $r = .367$, $p < .001$, Kappa = .341) for an accuracy of 71.9%. In the LOOCV, the DFA correctly allocated 194 of the 278 essays for an accuracy of 69.8%.

When the outcomes from the DFA models were compared to the actual scores assigned to the essays by the human raters, 120 of the 278 essays reported exact matches in scores (43% exact accuracy) and 226 of the essays reported adjacent matches in scores (81% adjacent accuracy). The confusion matrix provided in Table 13 indicates that for the total set of essays, the 104 essays in this set scored as a 1, 2, or 3 by humans were assigned a score of 4 or 5 by the algorithm. These errors arise from the earlier levels of the hierarchy because only scores of 4 and 5 are possible at this juncture. Nonetheless, 174 essays that were given a score of 4–6 by the human raters were allocated scores of either 4 or 5 by the algorithm (see Table 13).

3.3.4. Overall accuracy of scoring models for longer essays

In summary, the models developed for the essays that did meet the length threshold reported exact matches between the predicted essay scores and the human scores for 306 out of the 623 of the essays (49% exact accuracy; $df = 15$, $n = 623$, $\chi^2 = 280.596$, $p < .001$, $r = .538$, $p < .001$, Kappa = .419). The models reported adjacent matches for 552 of the essays (89% adjacent accuracy).

3.3.5. Overall accuracy for the entire data set

The models developed to score all the essays in our data set ($N = 1243$) reported exact matches between the predicted essay scores and the human scores for 689 out of the 1243 of the essays (55%

Table 13
Confusion matrix for the total set of longer essays showing actual and predicted essay scores.

Actual essay score	Predicted essay score	
	4	5
1	3	0
2	22	8
3	53	18
4	80	35
5	16	40
6	1	2

Table 14
Confusion matrix for the total set of essays showing actual and predicted essay scores.

Essay score	Predicted essay score					
Total set	1	2	3	4	5	6
1	**137**	53	33	3	0	0
2	37	**207**	116	22	8	0
3	2	69	**225**	53	18	0
4	0	10	71	**80**	35	0
5	0	1	4	16	**40**	0
6	0	0	0	1	2	0

exact accuracy; $df = 20$, $n = 1223$, $\chi^2 = 1175.775$, $p < .001$, $r = .714$, $p < .001$, Kappa = .566). The models reported adjacent matches for 1141 of the essays (92% adjacent accuracy; see Table 14 for the overall confusion matrix). The confusion matrix in Table 14 shows that the essay scores were overestimated for 341 essays and underestimated for 213 essays. This difference is partially due to the fact that there were many essays given a score of 1 ($n = 226$), which cannot be underestimated by the algorithm. The other source of the difference is from the tendency of the algorithm to overestimate those essays given a 2 by the raters: 116 (30%) were scored as a 3 by the algorithm.

4. Discussion

The purpose of this study was to demonstrate and evaluate the use of a hierarchical classification approach to providing automated assessment of essays. Our corpus included 1243 relatively heterogeneous essays from a variety of persuasive prompts and student grade levels. The essays were first divided in terms of the number of words and paragraphs, resulting in one set of shorter ($n = 620$) and one set of longer ($n = 623$) essays. As illustrated in Fig. 1, separate discriminant function analyses were then conducted, iteratively predicting the essay scores. This approach can be contrasted with one in which an algorithm uses a single set of variables to generate the prediction for an essay score. Our approach is similar to an iterative threshold-based approach in which a set of thresholds is used to generate sequential feedback. For example, the first round of feedback might regard the length of an essay, and the second, relevance, and so on. However, here the multiple iterations are also combined to generate a prediction regarding the overall quality of the essay.

For each iteration (or level within the hierarchy), we calculated the accuracy of the model's predictions. For the shorter essays, the overall accuracy was relatively high (62% exact accuracy; 95% adjacent accuracy), whereas it was somewhat lower accuracy for the longer essays (49% exact accuracy; 87% adjacent accuracy). Overall, the exact accuracy was 55% and the adjacent accuracy was 92%. These results indicate that a hierarchical approach can produce accuracy results comparable to those reported from other AES systems, such as e-rater (Attali & Burstein, 2006; Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012), IntelliMetric (Rudner, Garcia, & Welch, 2005; Rudner et al., 2006), and the Writing Pal (W-Pal; McNamara et al., 2013).

One goal in our writing research has been to extend current approaches to AES by developing new indices of essay quality (e.g., Crossley et al., 2012; Crossley, Defore, Kyle, Dai, & McNamara, 2013; Crossley & McNamara, 2010; McNamara et al., 2013) and to develop algorithms that will potentially facilitate and enhance feedback provided to students (Roscoe, Snow, & McNamara, 2013; Roscoe, Varner, Cai, Weston, Crossley, & McNamara, 2011; Roscoe, Varner (Allen), Crossley, & McNamara, 2013). This study continues this line of research by evaluating the potential of a hierarchical approach to drive feedback.

A guiding assumption in this study was that the quality of an essay at one level is not determined by the same indices as the quality of an essay at another level. The findings here confirm that different indices and different types of indices will inform each of the levels in the hierarchical analysis. For example, the indices that predicted the first level of differences among the shorter essays (see Table 4) were mainly related to the use of more sophisticated vocabulary and grammar (e.g., Flesch

Kincaid grade level, nominalizations, plural nouns, sentence relative clauses, commas), higher cohesion (lexical diversity, paragraph-to-paragraph), and the absence of references to more mundane, familiar topics such as social processes and religion. While the indices that were predictive of differences for longer essays (see Table 8) had some overlap (e.g., nominalizations, sentence relative clauses), the algorithm also included indices more reflective of the semantic content of the essay (e.g., lower bigram frequency, more frequent use of amplifiers and adjectives) and the use of more sophisticated verbs (i.e., less frequent use of public verbs, private verbs, verbs in base form, and auxiliary verbs).

These differences in algorithms confirm that a hierarchical approach can potentially yield layered features from which to develop formative feedback algorithms. Feedback can be based on linguistic features that focus on different attributes as a function of the relative quality of the essay. In contrast to previous models of essay quality (see McNamara, Crossley, et al., 2010), the type of feedback that can be provided is likely to be more salient in that the linguistic features assessed can lend themselves to practical interpretation. For example, in the case of an essay scored as a 2, after providing feedback on length and structure, suggestions could further be offered concerning specific aspects such as increasing *the amount of information* (e.g., increase nominalization, increase plural nouns, increase lexical diversity), increasing the *formality of writing* (e.g., decrease references to social processes, increase the use of third person pronouns, decrease the use of religious terms, decrease the use of public verbs, decrease the use of present participle verb forms, decrease narrativity), or increasing *semantic overlap* between paragraphs. This type of feedback contrasts with that geared toward an essay scored as a 4, which might focus more on strategies to improve parts of the essay (e.g., the introduction) or particular word choices (e.g., increase the use of amplifiers). The features at each level can be included in a series of feedback algorithms that could be either selected by the user or the importance (e.g., weights) within the scoring algorithm. Importantly, though, they lend themselves to simple instructions (e.g., consider showing less commitment to your ideas by including verbs such as *seem* and *appear* in place of verbs such as *be*) that can be practically interpreted by the writer to help revise and improve essay quality. It is beyond this paper to translate the specific indices at each level of the hierarchy into specific feedback; yet the point can be made that this approach affords gleaning information about specific features that may better inform feedback.

The particular indices included in this study are another aspect of our approach that may strengthen the impact of feedback. While many AES algorithms are driven principally by lower-level features, such as those related to grammar and spelling, we have focused on indices related to the linguistic, cognitive, and rhetorical features of the essays. On the one hand, it may seem that we throw the kitchen sink in terms of our choice of indices to include in the algorithm. On the other hand, the indices included in our tools are not arbitrary; they are theoretically motivated. The indices included in Coh-Metrix are guided by theories of how readers comprehend text and how linguistic features of text are related to text difficulty (e.g., Graesser & McNamara, 2011). The indices in LIWC are inspired by the assumption that particular discourse features can provide information about emotion and cognition. And, *WAT* is being developed on the heels of Coh-Metrix following the assumption that the linguistic features related to writing quality are different from those related to text difficulty, and other variables must be considered such as rhetorical cues and various semantics aspects of the writing. One of our goals has been to include indices related to various facets of discourse and cognition in order to more effectively inform the feedback that is provided to students.

While the goal of this study was to examine the potential advantages of using a hierarchical approach, several disadvantages also became apparent. One disadvantage is the increased complexity of the analysis. This analysis requires a series of computations that may include a large number of indices. This complexity may demand substantial computational resources and subsequently increase processing time. Given the power of current computer systems, this consideration may not present a substantial concern; however, it may result in problems in particular contexts or situations (e.g., web-based systems with many simultaneous users).

A second disadvantage is the fallout of errors from the top levels of the hierarchy. If the algorithm fails at a higher level, then subsequent levels cannot, as it stands, correct the error. For example, there were 33 essays that received a score of 1 or 2 by human raters but landed in the pool of higher relative quality essays that had met the length threshold (Group B2). These essays cannot be accurately

classified at this level of the hierarchy because only scores of 4 or 5 are possible at this level. Our (albeit optimistic) presumption was that these errors would be offset by the higher precision at each level; yet this may not always be the case. We also assume that these errors, while localized to higher scores, may be rare because essay scores in the higher quality bins are somewhat rarer than essays in the lower bins. Nevertheless, feedback based on this hierarchical approach is sensitive to the assigned quality level of the essay and thus future studies might consider a using a more stratified corpus to develop scoring algorithms (if available).

It is clear that this study is only a starting point and there are many potential studies that might follow from it or improve upon it. One important consideration regards the choice in machine learning analytic techniques. In this study, we used DFA to generate the predictions at each level of the hierarchy. DFA is one of many machine-learning techniques that can be used for these types of analyses. Others include naïve Bayes classifiers, decision tree classifiers, and support vector machines. When conducting this study, we began with the assumption that each prediction within the hierarchy *need not* use the same type of analysis. For example, an algorithm rather than a threshold might be used to implement the initial division and different analytical techniques might be used to predict lower and higher quality responses at each level. However, our initial analyses did not yield clear advantages for other machine learning techniques. Hence, the initial length threshold and DFA were the statistical techniques we adopted for this study. Future studies might consider the utility of using other statistical techniques. Different machine learning approaches may be more or less useful depending on the type of data. In addition, the combination of a variety of statistical approaches across different levels of a hierarchy may be more effective: different levels of quality may call upon the use of different techniques. The important contribution of this study is to demonstrate the potential value of this technique and point toward the *possibility* of using different techniques at different levels of an analysis.

An additional consideration is that this study focuses on the accuracy of the algorithm in relation to expert ratings, and not on the algorithm's validity in driving feedback that affects improvement in essay quality (e.g., McNamara et al., 2013; Roscoe, Snow, et al., 2013; Roscoe, Varner (Allen), et al., 2013). Ideally, a study is needed in which one group of students is provided with feedback driven by the hierarchical algorithm and another group of students is provided with feedback driven by a contrasting algorithm, such as an algorithm that focuses on lower-level traits. Such a study is certainly needed to make strong conclusions; but currently, this is left for future research.

Notably, this research presupposes a need for and general acceptance for automated scoring of essays. From our standpoint, we see this need as self-evident. If students are to be provided with sufficient practice and feedback on writing to improve in writing skills, and if students are to be academically judged in terms of these writing skills on a large scale (e.g., the Test of English as a Foreign Language and the Graduate Record Exam), then it is necessary to relieve the subsequent burden on teachers and raters by developing increasingly accurate and valid scoring techniques. Nonetheless, we recognize that our own viewpoint does not mirror that of some educators and researchers who emphatically oppose the use of AES (e.g., Cheville, 2004; Ericsson & Haswell, 2006; Herrington & Stanley, 2012; Jones, 2006; McGee, 2006; Perelman, 2012). One marked objection regards the social nature of writing, and the assumed inability of automated scoring techniques to capture these social nuances. Indeed, many subtleties and variability in factors such as knowledge, cultural and linguistic background, and the rhetorical purpose of the writing can make substantial differences in text quality and characteristics (e.g., Beck & Jeffery, 2007; McNamara, 2013; Murphy & Yancey, 2008). However, rather than view these factors as game-stoppers to AES, we prefer to view these as challenges to acknowledge and eventually overcome. Indeed, one important limitation of AES in general is that algorithms do not consider factors beyond the features of the writing sample itself, such as the writer's prior performance, prior literacy skills, prior knowledge of the targeted domain, epistemic frame, native language, and so on. Such considerations would bring AES closer to the intelligent tutoring approach by modifying algorithms based on a student model (e.g., VanLehn, 1988). This limitation is primarily a consequence of how AES research is conducted, mostly in the absence of information about the writer, at least when large samples of writing are collected. Perhaps a next step in AES research is to move toward tailoring algorithms according to individual differences, or at least to examine the benefits of doing so.

5. Conclusion

Our exploration of the hierarchical approach is intended to expand the array of methods that might be used when developing AES systems (see Elliot & Klobucar, 2013, for the role of innovation in AES systems). The results indicate that the accuracy of the scores is equivalent to accuracy reported using techniques in which the features of the essay are considered simultaneously in the algorithm (Attali & Burstein, 2006; Ramineni et al., 2012; Rudner et al., 2005, 2006). Although there are some potential disadvantages to using a hierarchical approach, one notable advantage is the potential to better inform feedback to the writer. If the goal of a system is to provide formative feedback, then the use of different variables at different levels may have greater potential to be useful. As such, the implementation of this hierarchical approach within a tutoring system may provide benefits well beyond those provided by scoring models based on more simple regression analyses. Indeed, this approach to automated evaluation might be applied to any number of issues and problems. Here we have focused on the scoring of persuasive essays. Assumedly, however, the approach might be applied to other natural language problems, such as the scoring of short answers to comprehension or knowledge assessments. The features that inform the assessments will be different depending on the particular problem at hand, yet this approach has the potential to yield more accurate and informative performance models than has simple one-shot regression.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grants R305A120707 and R305A090623 to Arizona State University. The opinions expressed are those of the author and do not represent views of the Institute or the US Department of Education.

Appendix. SAT scoring rubric[2]

"The essay will be scored by experienced and trained high school and college teachers. Each essay will be scored by two people who won't know each other's score. They won't know the student's identity or school either. Each reader will give the essay a score from 1 to 6 (6 is the highest score) based on the following scoring guide."

SCORE OF 6: Demonstrates clear and consistent mastery, although it may have a few minor errors. Effectively and insightfully develops a point of view on the issue and demonstrates outstanding critical thinking, using clearly appropriate examples, reasons, and other evidence to support its position is well organized and clearly focused, demonstrating clear coherence and smooth progression of ideas exhibits skillful use of language, using a varied, accurate, and apt vocabulary, meaningful variety in sentence structure, free of most errors in grammar, usage, and mechanics.

SCORE OF 5: Demonstrates reasonably consistent mastery, although it will have occasional errors or lapses in quality. A typical essay effectively develops a point of view on the issue and demonstrates strong critical thinking, generally using appropriate examples, reasons, and other evidence to support its position is well organized and focused, demonstrating coherence and progression of ideas exhibits facility in the use of language, using appropriate vocabulary, variety in sentence structure, generally free of most errors in grammar, usage, and mechanics.

SCORE OF 4: Demonstrates adequate mastery, although it will have lapses in quality. A typical essay develops a point of view on the issue and demonstrates competent critical thinking, using adequate examples, reasons, and other evidence to support its position is generally organized and focused, demonstrating some coherence and progression of ideas exhibits adequate but inconsistent facility in the use of language, using generally appropriate vocabulary demonstrates some variety in sentence structure has some errors in grammar, usage, and mechanics

[2] Excerpted from: http://www.collegeboard.com/student/testing/sat/about/sat/essay_scoring.html.

SCORE OF 3: Demonstrates developing mastery, and is marked by ONE OR MORE of the following weaknesses: develops a point of view on the issue, demonstrating some critical thinking, but may do so inconsistently or use inadequate examples, reasons, or other evidence to support its position is limited in its organization or focus, or may demonstrate some lapses in coherence or progression of ideas displays developing facility in the use of language, but sometimes uses weak vocabulary or inappropriate word choice lacks variety or demonstrates problems in sentence structure contains an accumulation of errors in grammar, usage, and mechanics.

SCORE OF 2: Demonstrates little mastery, and is flawed by ONE OR MORE of the following weaknesses: develops a point of view on the issue that is vague or seriously limited, and weak critical thinking, providing inappropriate or insufficient examples, reasons, or other evidence to support its position is poorly organized and/or focused, or demonstrates serious problems with coherence or progression of ideas displays very little facility in the use of language, using very limited vocabulary or incorrect word choice demonstrates frequent problems in sentence structure contains errors in grammar, usage, and mechanics so serious that meaning is somewhat obscured.

SCORE OF 1: Demonstrates very little or no mastery, and is severely flawed by ONE OR MORE of the following weaknesses: develops no viable point of view on the issue, or provides little or no evidence to support its position is disorganized or unfocused, resulting in a disjointed or incoherent essay displays fundamental errors in vocabulary demonstrates severe flaws in sentence structure contains pervasive errors in grammar, usage, or mechanics that persistently interfere with meaning Essays not written on the essay assignment will receive a score of zero.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*, *4*(3). Retrieved from www.jtla.org

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Beck, S. W., & Jeffery, J. (2007). Genres of high stakes writing assessments and the construct of writing competence. *Assessing Writing*, *12*, 60–79.

Bianchi, N., Gentile, C., & Zaniboni, L. (2006). Incremental algorithms for hierarchical classification. *The Journal of Machine Learning Research*, *7*, 31–54.

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

Bridgeman, B. (2013). Human ratings and automated essay evaluation. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay scoring: Current applications and future directions* (pp. 221–232). New York: Routledge.

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 perspective* (pp. 113–122). Hillsdale, NJ: Lawrence Erlbaum.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing system. *AI Magazine*, *25*, 27–36.

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

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

College Board. (2011). *Essay scoring guide: A framework for scoring SAT essays*. Retrieved from http://professionals.collegeboard.com/testing/satreasoning/scores/essay/guide

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *2*, 213–238.

Crismore, A., Markkanen, R., & Steffensen, M. S. (1993). Metadiscourse in persuasive writing: A study of texts written by American and finish university students. *Written Communication*, *39*, 39–71.

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., Defore, C., Kyle, K., Dai, J., & McNamara, D. S. (2013). Paragraph specific n-gram approaches to automatically assessing essay quality. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 216–219). Heidelberg, Berlin, Germany: Springer.

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984–989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1231–1236). Austin, TX: Cognitive Science Society.

Crossley, S. A., Roscoe, R., McNamara, D. S., & Graesser, A. C. (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). Auckland, New Zealand: AIED.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*, 573–605.
Crossley, S. A., Weston, J. L., 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*, 282–311.
Crossley, S. A., White, M. J., McCarthy, P. M., & McNamara, D. S. (2009, November). The effects of elaboration and cohesion on human evaluations of writing proficiency. In *Poster presented at the 39th annual meeting of the Society for Computers in Psychology* Boston, MA.
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. *Journal of Technology, Learning and Assessment*, 5. Retrieved from http://www.jtla.org
Douglas, D. (1981). An exploratory study of bilingual reading proficiency. In S. Hudelson (Ed.), *Learning to read in different languages* (pp. 33–102). Washington, DC: Center for Applied Linguistics.
Dufty, D. F., Graesser, A. C., Lightman, E., Crossley, S. A., & McNamara, D. S. (2006). An algorithm for detecting spatial cohesion in text. In *Paper presented at the 16th annual meeting of the Society for Text and Discourse* Minneapolis, MN.
Dufty, D. F., Graesser, A. C., Louwerse, M., & McNamara, D. S. (2006). Assigning grade level to textbooks: Is it just readability? In R. Sun, & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the Cognitive Science Society* (pp. 1251–1256). Austin, TX: Cognitive Science Society.
Dufty, D., Hempelmann, C., Graesser, A., Cai, C., & McNamara, D. S. (2005). An algorithm for detecting causal and intentional information in text. In *Paper presented at the 15th annual meeting of the Society for Text and Discourse* Amsterdam.
Dumais, S. T., & Chen, H. (2000). Hierarchical classification of web content. In *Proceedings of the 23rd ACM international conference on research and development in information retrieval* (pp. 256–263). ACM Press.
Elliott, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis, & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Erlbaum.
Elliot, N., & Klobucar, A. (2013). Automated essay evaluation and the teaching of writing. In M. D. Shermis, J. Burstein, & S. Apel (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 16–35). London: Routledge.
Ericsson, P. F., & Haswell, R. (Eds.). (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.
Fearn, L., & Farnan, N. (2005, April). An investigation of the influence of teaching grammar in writing to accomplish an influence on writing. In *Paper presented at the annual meeting of the American Educational Research Association* Montreal, Canada.
Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press (CD-ROM).
Ferris, D. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, *28*, 14–20.
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 scoring: Current applications and future directions* (pp. 68–88). New York: Routledge.
Frase, L., Faletti, J., Ginther, A., & Grant, L. (1997). *Computer analysis of the TOEFL test of written English (TOEFL research rep. no. 64)*. Princeton, NJ: Educational Testing Service.
Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *2*, 371–398.
Graesser, A. C., & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In H. Cooper, P. Camic, R. Gonzalez, D. Long, & A. Panter (Eds.), *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics*. Washington, DC: American Psychological Association.
Graesser, A. C., McNamara, D. S., & Kulikowich, J. (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*, 445–476.
Granitzer, M. (2003). *Hierarchical text classification using methods from machine learning* (Unpublished doctoral dissertation). Austria: Graz University of Technology.
Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, *8*, 4–43.
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.
Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
Herrington, A., & Stanley, S. (2012). Criterion℠: Promoting the standard. In A. B. Inoue, & M. Poe (Eds.), *Race and writing assessment* (pp. 47–61). New York, NY: Peter Lang.
Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*, 237–263.
Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, *47*, 549–566.
Jarvis, S. (2011). Data mining with learner corpora: Choosing classifiers for L1 detection. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora. In honour of Sylviane Granger* (pp. 127–154). Amsterdam: John Benjamins.
Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*, 377–403.
Jones, E. (2006). ACCUPLACER'S essay-scoring technology: When reliability does not equal validity. In P. F. Ericsson, & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 93–113). Logan, UT: Utah State University Press.
Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kellogg, R., Whiteford, A., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research*, *42*, 173–196.
Kincaid, J., Fishburne, R., Rogers, R., & Chissom, B. (1975). *Derivation of new readability formulas for navy enlisted personnel. Branch Report 8-75*. Millington, TN: Chief of Naval Training.
Kintsch, W., & van Djik, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363–394.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.
Landauer, T., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education*, *10*, 295–308.
Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.
Longo, B. (1994). The role of metadiscourse in persuasion. *Technical Communication*, *41*, 348–352.
Louwerse, M. M. (2001). *An analytic and cognitive parameterization of coherence relations*. *Cognitive linguistics* (Vol. 12) Mahwah, NJ: Erlbaum.
Malvern, D. D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke, UK: Palgrave Macmillan.
McGee, T. (2006). Taking a spin on the Intelligent Essay Assessor. In P. F. Ericsson, & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 79–92). Logan, UT: Utah State University Press.
McNamara, D. S. (2013). The epistemic stance between the author and the reader: A driving force in the cohesion of text and writing. *Discourse Studies*.
McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*, 57–86.
McNamara, D. S., Crossley, S. A., & Roscoe, R. D. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*, 499–515.
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., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.
McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*, 292–330.
Meadows, M., & Billington, L. (2005). *Review of the literature on marking reliability. Report for the Qualifications and Curriculum Authority*. London: National Assessment Agency.
Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, *3*, 235–244.
Murphy, S., & Yancey, K. B. (2008). Construct and consequence: Validity in writing assessment. In C. Bazerman (Ed.), *Handbook of writing research: History, society, school, individual, text* (pp. 365–386). New York, NY: Lawrence Erlbaum.
Pearson, P. D. (1974–1975). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relationships. *Reading Research Quarterly*, *10*, 155–192.
Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count: LIWC 2007*. Austin, TX: LIWC.net. www.liwc.net
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/Anderson, SC: WAC Clearinghouse/Parlor Press.
Raine, R. B., Mintz, L., Crossley, S. A., Dai, J., & McNamara, D. S. (2011). Text box size, skill, and iterative practice in a writing task. In R. C. Murray, & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference (p* (pp. 537–542). Menlo Park, CA: AAAI Press.
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.
Ramineni, C., & Williamson, D. M. W. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*, 25–39.
Rashotte, C. A., & Torgesen, J. K. (1985). Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly*, *20*, 180–188.
Roscoe, R. D., Snow, E. L., & McNamara, D. S. (2013). Feedback and revising in an intelligent tutoring system for writing strategies. In K. Yacef, K. Yacef, et al. (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED)* (pp. 259–268). Heidelberg, Berlin: Springer.
Roscoe, R. D., Varner, L. K., Cai, Z., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2011). Internal usability testing of automated essay feedback in an intelligent writing tutor. In R. C. Murray, & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 543–548). Menlo Park, CA: AAAI Press.
Roscoe, R. D., Varner (Allen), L. K., Crossley, S. A., & McNamara, D. S. (2013). Developing pedagogically-guided threshold algorithms for intelligent automated essay feedback. *International Journal of Learning Technology*, *8*, 362–381.
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)*. Retrieved from http://www.gmac.com/gmac/researchandtrends/
Rudner, L., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *Journal of Technology Learning, and Assessment*, *4*, 3–21.
Schultz, M. T. (2013). The IntelliMetrix automated essay scoring engine—A review and an application to Chinese essay scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay scoring: Current applications and future directions* (pp. 89–98). New York: Routledge.
Shermis, M., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Shermis, M., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York: Routledge.
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*. Oxford, UK: Elsevier.
Shermis, M., Koch, C., Page, E., Keith, T., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement*, *62*, 5–18.
Streeter, L., Psotka, J., Laham, D., & MacCuish, D. (2002). The credible grading machine: Essay scoring in the DOD. In *The interservice/industry training, simulation & education conference (I/ITSEC)*.
Templin, M. (1957). *Certain language skills in children: Their development and interrelationships*. Minneapolis, MN: The University of Minnesota Press.
Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, *2*, 319–330.
VanLehn, K. (1988). Student modeling. In M. Polson, & J. Richardson (Eds.), *Foundations of intelligent tutoring systems*. Hillsdale, NJ: Erlbaum.
Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*, 22–36.
Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, *10*, 1–24.
Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *1*, 85–99.
Weston, J. L., Roscoe, R., Floyd, R. G., & McNamara, D. S. (2013, April). The WASSI (Writing Attitudes and Strategies Self-Report Inventory): Reliability and validity of a new self-report writing inventory. In *Poster presented at the 2013 annual meeting of the American Educational Research Association* San Francisco, CA.
Wilson, M. D. (1988). The MRC psycholinguistic database: Machine-readable dictionary (2nd version). *Behavioral Research Methods, Instruments and Computers*, *20*, 6–11.
Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the myside bias in written argumentation. *Written Communication*, *26*, 183–209.
Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*, 291–300.

Danielle S. McNamara is a Professor of Psychology at Arizona State University. She focuses on educational technologies and discovering new methods to improve students' ability to understand challenging text and convey their thoughts and ideas in writing. Her work (see http://soletlab.com) integrates various approaches including the development of game-based tutoring systems (e.g., iSTART, Writing Pal), the development of natural language processing tools, and the use of learning analytics across multiple contexts.

Scott Crossley is an Associate Professor at Georgia State University. His interests include computational linguistics, corpus linguistics, cognitive science, discourse processing, and discourse analysis. His primary research focuses on corpus linguistics and the application of computational tools in second language learning and text comprehensibility.

Rod D. Roscoe is an Assistant Professor in the Department of Human and Environmental Systems, Cognitive Science and Engineering Program, at Arizona State University. His research focuses on self-regulated learning processes in formal and informal contexts and examines how these processes can be facilitated via adaptive technology, instruction, and peer support.

Laura K. Allen is a Ph.D. student in Psychology and the Learning Sciences Institute at Arizona State University. Her research examines the cognitive processes and abilities underlying reading and writing proficiency, and also considers the impact of these factors in second language learning.

Jianmin Dai is a Research Assistant Professor in the Learning Science Institute at Arizona State University. He has contributed to design and development of several systems including iSTART, the Writing Pal, and Coh-Metrix T.E.R.A. His research focuses on the application of natural language processing and machine learning in game-based and intelligent tutoring systems.