# What automated analyses of corpora can tell us about students' writing skills

Paul Deane° & Thomas Quinlan*

° Educational Testing Service, Princeton, New Jersey | USA

* The College of New Jersey, Ewing, New Jersey | USA

**Abstract:** A particular application of corpus analysis, automated essay scoring (AES) can reveal much about students' writing skills. In this article we present research undertaken at Educational Testing Service (ETS) as part of its ongoing commitment to developing effective AES systems. AES systems have certain advantages. They can: (a) produce scores similar to those assigned trained human raters, (b) provide a single consistent metric for scoring, and (c) automate linguistic analyses. However, to understand student writing, we may need to look beyond the final essay in various ways, to consider both the process and the product. By broadening our definition of corpora, to capture the dynamics of written composition, it may become possible to identify profiles of writing behavior.

**Keywords**: automated essay scoring, corpus analysis, writing assessment

journal of
WRITING RESEARCH

Deane, P., & Quinlan, T. (2010).What automated analyses of corpora can tell us about students' writing skills *Journal of Writing Research, 2* (2), 151-177. http://dx.doi.org/10.17239/jowr-2010.02.02.4

## 1. What Automated Analyses of Corpora Can Tell Us About Students' Writing Skills

Writing skill is crucial to students' success in school and beyond, making it an important focus of instruction and assessment. While all academic assessments require careful application of the principles of test design (Crocker & Algina, 1986), writing assessment poses unique challenges—while also presenting opportunities. In this article we will present research that has been undertaken at Educational Testing Service (ETS) as part of its ongoing commitment to developing effective Automated Essay Scoring (or AES) systems, in particular e-rater®. This research, based upon analyses of extensive corpora of scored student writing, provides a detailed example of how corpus data can be applied to the analysis of writing, and how it can support ongoing efforts to build models of writing skill and writing development. While the immediate application of an AES model is practical, supporting large-scale operational assessments, this work illustrates how corpus-based techniques can inform cognitive approaches to understanding the development of writing skills.

## 2. General Background to AES

Automated essay scoring systems (AES) can consistently capture machine-detectable features germane to writing quality. Such systems are developed by analyzing large corpora of student essays, first to identify useful features and then to build scoring models in which human ratings of essay quality are used as an external criterion. The primary limitation of such systems is the nature of the features that can be measured, which depends in turn upon the application of natural language processing (NLP) techniques. Such techniques are applied most easily to mechanical features--such as grammar and spelling, or even the organizational patterns of text--and are difficult to apply to deeper aspects of writing, such as quality of argumentation. However, there are strong correlations across all features that define skillfully written text, at least as assessed by human raters (Godshalk, College Entrance Examination Board, & et al., 1966). Human holistic scores thus provide a reasonably reliable measure of general text quality, which can be used in turn to train an automated model. Such methods have obvious applications to the operational scoring of writing assessments, but can also be leveraged to provide detailed information how writing skills may develop over time.

AES systems can be traced back to Project Essay Grade (Page, 1966; Page, 2003), which used a variety of simple textual features as proxies for writing quality. More recently a variety of AES methods have been developed. For example, Intelligent Essay Assessor™ applies (among others) one NLP technique, Latent Semantic Analysis (Foltz, Kintsch, & Landauer, 1998; Foltz, Laham, & Landauer, 1999a; Foltz, Laham, & Landauer, 1999b; Landauer & Dumais, 1997), to the problem of scoring essays. The e-rater® scoring engine relies upon a variety of NLP techniques to detect mechanical, grammatical, lexical and discourse features of student essays (Yigal, Attali & Burstein,

2006a; Burstein, Kukich, Wolff, Lu, & Chodorow, 1998). These include spellchecking, methods that detect characteristic word sequences (also known as n-grams) that indicate the presence of grammatical errors, and the application of a discourse parser to recognize key structural elements of an essay. In addition, for some applications, the e-rater scoring engine includes two features that rely upon Content Vector Analysis (Salton, Yang, & Wong, 1975), which is a close relative of LSA.

AES can be viewed as an instance of a more general research tradition that uses text features to measure properties of interest, such as essay quality, text readability, genre characteristics, or linguistic mastery. Research into t-unit length provides an apt example (Hunt, 1970). In writing research, t-unit[1] length has often been used to measure the development of syntactic complexity. Studies of the relation between t-unit length and holistic writing quality have yielded inconsistent results, at least for L1 writers. T-unit length significantly predicted quality: only for grade 5 writers, not for grades 8 or 11 (Stewart & Grobe, 1979); only for lower quality essays (Witte, Daly, & Cherry, 1986); only for grade 10, not for grades 6 or 12 (Crowhurst, 1980). For L2 writers, the evidence seems clearer. Cumming and his colleagues (2006) analyzed essays composed by adult EFL writers, and found that higher-quality essays had significantly longer t-units than lower-quality essays. These results suggest that syntactic complexity (as measured by t-unit length) may be more meaningful as a proxy of writing skill in some populations and contexts than it is in others. The moral can be generalized: various text features are implicated in writing quality, but to different degrees for different populations. The practical implication is that operational tests may require scoring models specifically tuned by prompt or population. However, there is also a theoretical implication: Corpus analysis of textual features, using AES methods, can help map out the dimensions of writing quality, and how they change developmentally.

One aspect of writing skill is the ability to produce well-developed documents. Accordingly, much AES research focuses on features that arguably measure fluency, complexity, and accuracy in language production. Many of the same features are also implicated in readability metrics, which generally include sentence length and vocabulary frequency (Dale & Chall, 1948; Dale & Tyler, 1934; Flesch, 1948; Lively & Pressey, 1923; Ojemann, 1934; Patty & Painter, 1931; Vogel & Washburne, 1928). A more recent research tradition applies quantitative methods to genre analysis (Biber, 1988, 1995a, 1995b; Biber, Conrad, & Reppen, 1998). In this approach, a variety of features are extracted, and their covariance across a large corpus is investigated by factor analysis. Genres are then defined in terms of clusters of linguistic features typically shared across texts written for similar purposes and audiences. There appear to be strong connections between the kinds of variation in written text that reflect genre variation and those that reflect variations in readability, and writing quality, as will be discussed in more detail below.

Features used in AES should not just predict human judgments empirically, but also have a clear theoretical rationale. There is strong validity support for automated

measures of essay quality, the extent to which they (a) demonstrate a strong relationship to human qualitative judgments and (b) reflect known developmental trends. For example, writing skills generally improve with age, so automated measures should reflect developmental progressions in which lexical and grammatical complexity increase with age and maturity (Loban, 1976). Similarly, among foreign language learners, a developmental pattern has been observed in which increasing proficiency can be measured using features that reflect the fluency of text production, the accurate use of lexical and grammatical patterns, and the richness and syntactic complexity of spontaneously produced texts (Wolfe-Quintero, Inagaki, & Kim, 1998).

Conversely, validated AES techniques can provide an important method for corpus analysis, enabling analysis of much larger corpora, at a finer level of detail, than would be possible if all corpus features had to be annotated by hand. For several years, ETS researchers have been developing and testing corpus-based methods for automatically scoring student writing. As AES techniques mature, it becomes possible to draw some preliminary conclusions about the potential of automated corpus analysis as a means for understanding the development of writing skill.

## 3. Predicting Human Judgments of Essay Quality

The validity of essay examinations rest upon an argument: that writing skill can be measured by examining the quality of essays written in response to (one or more) impromptu topics. Thus, essay examinations require: (a) defining a reasonable sample by which to judge writing skill, (b) defining defensible standards for measuring the writing quality of each sample, and then (c) applying those standards consistently (Breland, Camp, Jones, Morris, & Rock, 1987; Huot, 1993).

Defining a reasonable sample depends in part upon what aspects of writing quality a test is intended to measure. If the primary goal is to measure fundamental fluency and accuracy of text production, relatively small samples gathered on one or two occasions of use may suffice. As the construct is expanded to include higher-order writing skills, including the ability to handle different styles and genres of writing and the ability to address different audiences and topics, much larger samples may be necessary, leading in some cases to a preference for portfolio assessments (Huot, 1996).

Although the notion of essay "quality" poses some thorny psychometric issues, there is considerable consensus about many essential aspects of "quality." For example, the 6-Trait scoring approach was developed by group of language arts experts who were tasked with identifying essential aspects of essay quality (Spandel & Stiggins, 1990). The traits they identified seem to recur whenever educators are asked to define essay quality: substantive, interesting ideas and content, a distinctive voice, fluency and clarity of expression, effective word choice, and adherence to conventions[2]. While it is fairly easy to define relevant standards, and thus to establish scoring criteria and develop rubrics, it is far more difficult to apply such scoring criteria in a consistent fashion.

In the history of assessing writing, interrater reliability has been an ongoing focus of concern (Elliot, 2005). Considerable research indicates that rater judgments of essay quality tend to lack stability, when based upon the judgment of a single individual (Breland et al., 1987; Diederich, 1974; Diederich, French, & Carlton, 1961). Holistic scoring techniques were developed to improve the stability of judgment of essay quality (Diederich, 1974; Diederich et al., 1961).

Holistic scoring techniques require raters to form an overall impression of essay quality, while taking into account all the scoring criteria. Holistic scoring is clearly a complex task, requiring raters to balance some criterion against others. The evidence suggests that trained raters can successfully apply the scoring criteria (Huot, 1993), although they sometimes differ widely in how they apply them, in terms of emphasizing some criteria over others (Lumley, 2002). This tendency to apply rubrics in somewhat idiosyncratic ways should seemingly lead raters scores to diverge. Yet, trained raters routinely achieve satisfactory levels of interrater agreement in operational scoring of large-scale writing assessments. How might we reconcile this apparent contradiction? Research evidence suggests that traits of essay quality are highly correlated (Lee, Gentile, & Kantor, 2008; McNamara, 1990), such that an essay strong on one trait (e.g., word choice) will often be strong on another (e.g., organization). Accordingly, in "forcing" essays onto a single scale, holistic scoring takes advantage of the natural covariance of quality traits.

The alternative is to score analytically, by trait. One such method is "6-trait" scoring. In collaboration with a group of classroom teachers, Vicki Spandel (1990) identified six dimensions of essay quality for the purposes of providing students with teacher feedback: ideas/content, organization, voice, word choice, sentence fluency, and conventions. Like other trait scoring methods, there are very high correlations across trait-scores (Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006). While the lack of separability among human trait scores is problematic, the traits identified by teachers do represent legitimate targets of writing instruction.

Automated essay scoring systems function on a similar principle. In composing a text, the writer must arrange letters into words, words into sentences, and sentences into paragraphs. Accordingly, it should not be surprising to discover that certain linguistic aspects of an essay relate strongly to rater judgments of overall essay quality. Some of these linguistic aspects can be detected by NLP capabilities. Of themselves, these linguistic aspects may cover only a limited portion of the construct. Generally, educators recognize that "writing skill" goes beyond basic skills (e.g., spelling, grammar, and punctuation), and very much includes abilities to select and organize ideas. In writing, simple things (i.e., words and phrases) are used to build complex things (i.e., documents), with the former to a great extent controlling the latter. So, the measurement of linguistic aspects can reveal quite a bit about essay quality. They may not capture some essential aspects of quality, such as ideas or voice, but they can measure many of the traits characteristic of stronger writers, including fluency, word choice, adherence to conventions, and use of appropriate discourse structures.

Linguistic measures can thus be useful for predicting essay quality without measuring all aspects of the construct directly.

ETS has applied NLP techniques to essay scoring in order to address a particular need. In large-scale writing assessments, essays are typically double-scored, by two trained human raters, using a holistic rubric (a scoring guide that guides the human rater in forming an overall impression of an essay's quality, by taking into account various dimensions of essay quality, without making an explicit enumeration of specific strengths, weaknesses, and errors). When raters are well-trained, guaranteeing a satisfactory level of inter-rater agreement, double-scoring yields reliable results. However, considerable effort must be expended on rater training, and the entire scoring process must be calibrated carefully so that different groups of raters, assessing different sets of essays under varying conditions, nevertheless apply the same, consistent standards. ETS researchers have therefore focused on developing an AES system that is at least as reliable as trained human raters, and thus can be used to reduce the need for double-scoring by providing a reliable, automated crosscheck on rater judgments.

The development of NLP capabilities proceeds as follows. Given a corpus of student essays, researchers consider a particular aspect of essay quality and write a computer program to capture it. The program is trained on one part of the corpus, then cross-validated against the other. The program output defines a NLP feature, and once that feature demonstrates an acceptable level of precision[3], it may be included in the essay scoring engine.

The e-rater® scoring engine developed at ETS employs features that measure aspects of grammar, usage, mechanics, style, organization, development, lexical complexity, and prompt-specific vocabulary usage (i.e. content) (Attali & Burstein, 2006). Most features (i.e., Grammar, Usage, Mechanics, Style, and Lexical Sophistication) are aggregated across multiple sub-measures, or microfeatures. Figure 1 illustrates the decomposition of e-rater into features and microfeatures.

The features and microfeatures have two applications. The aggregate features are used as inputs to the e-rater scoring engine, which is discussed in more detail below. Individual microfeatures are also used as part of a formative feedback system, Criterion®, which scores student essays and provides feedback to them about errors in grammar, usage, mechanics, style, vocabulary, organization and development (Burstein, Chodorow, & Leacock, 2003). The specific features presented in Figure 1 are those in actual operational use in e-rater and Criterion, and are the result of a long process of feature development based upon NLP analysis of corpus data, documented in a series of research reports and articles (Burstein, Andreyev, & Lu, 2002; Burstein & Higgins, 2005; Burstein, Kukich, Wolff, Lu, Chodorow et al., 1998; Burstein & Marcu, 2000, 2003; Burstein, Marcu, Andreyev, & Chodorow, 2001; Chodorow & Leacock, 2000; Higgins, Burstein, Marcu, & Gentile, 2004).

In e-rater, human ratings of writing quality are predicted from feature scores by calculating a weighted average. The weighting is achieved by one of two methods to create either prompt-specific or generic models. Model-building proceeds as follows.

ETS staff selects a corpus of essays scored by two trained human raters (i.e., double-scored essays), processes them through e-rater to obtain feature scores, and then uses regression analysis to determine what weighting scheme best predicts average human scores. With prompt-specific models, a distinct weighting scheme is determined for each writing task. With generic models, a regression model is determined for essays drawn from a group of prompts, typically defined by population, such as TOEFL test-takers, or 8th grade American middle-school students, though genre distinctions are typically also made, distinguishing (for instance) persuasive from expository prompts. E-rater scores (whether produced by prompt-specific or generic models) have shown to predict human holistic scores about as well as one human score predicts another (Yigal, Attali & Burstein, 2006b).

While e-rater scores strongly predict human scores, we can ask what this prediction signifies? To what extent do e-rater scores cover the construct of writing skill? Models of writing expertise posit the coordination of multiple cognitive processes, which compete for limited working memory resources (Flower & Hayes, 1981; Kellogg, 1996; McCutchen, 1996). In these models, a common assumption is that inefficient low-level processes (e.g., handwriting and spelling) can interfere with high-level processes (e.g., solving problems related to content and rhetoric). In particular, unless a writer can produce text with some degree of fluency, text production will draw cognitive resources away from the process of generating and organizing ideas. This process may account for an observation that holds consistently across a wide range of writing situations: in most timed writing tasks, there is a strong correlation between essay length and holistic ratings of essay quality (Powers, Burstein, Chodorow, Fowles, & Kukich, 2001).

On the one hand, e-rater explicitly values what might seem to be low-level constructs: grammar, usage, mechanics, style, among others, with relatively little representation of higher order cognitive skills. However, higher order processing depends upon the coordination of multiple low-level skills, which are directly responsible for text production. The validity of the e-rater scoring model thus relies upon the existence of strong correlations between various aspects of writing quality, even if it does not measure some aspects of writing quality explicitly. Interpreting these correlational relationships, we have reason to suppose that e-rater succeeds in measuring aspects of basic writing skill, which in turn provides strong prediction of student ability to apply a more critical approach to literacy. Framed in terms of Bereiter and Scardamalia's (1987) models of writing, we might say that e-rater scoring provides a measure of knowledge-telling, which can serve as a fairly reliable predictor of students' abilities to adopt a knowledge-transforming approach.

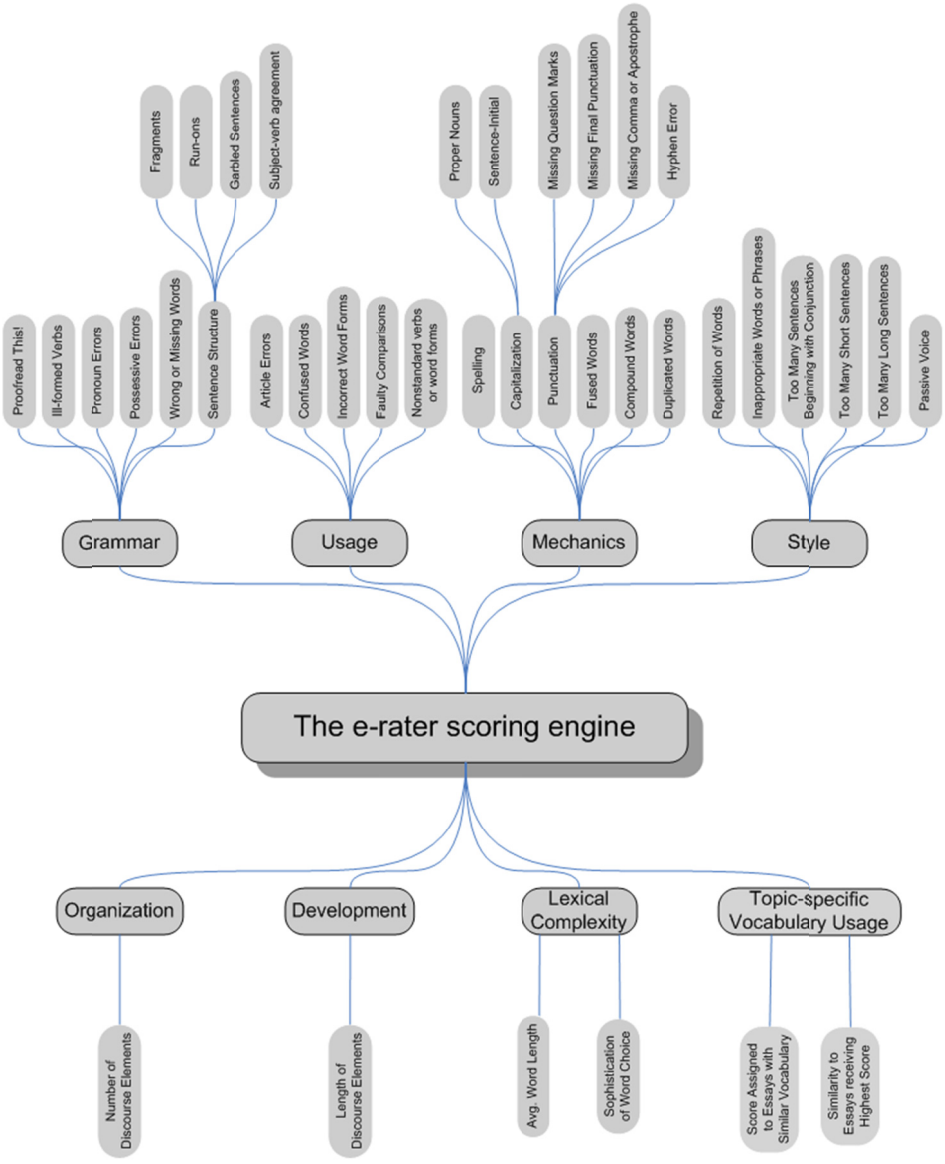


**Figure 1.** The decomposition of e-rater into features and microfeatures. Adapted from *Evaluating the Construct Coverage of the E-rater Scoring Engine*, by T. Quinlan, D. Higgins, & S. Wolff, 2009.

We recognize that predicting holistic quality scores can reveal only so much about examinees' writing skills. Educational leaders in the U.S. have asked that assessments provide information to support instruction and learning, and for that purpose prediction of holistic scores can provide crucial information, since it helps to identify those students who most require intervention, particularly those in need of support to improve the fluency and accuracy of text production. But the individual features that enter into the prediction of the holistic score can also be interpreted as partial measures of particular traits – perhaps not so reliably as to yield trustworthy individual scores, but well enough to characterize group differences. Thus, AES has the potential to be used to profile student performance across groups and contexts. With a large corpus of scored responses for a specific grade level or other population, a model can be constructed whose features define what high, medium, and low quality writing looks like. From a cognitive perspective, the set of models thus built comprise foundational data, since a good cognitive model ought to be able to make sense of how individuals' pattern of writing performance varies across prompts, under different writing conditions. As a result, AES methods can contribute heavily to an investigation of writing skill by providing strong corpus-defined anchors to clarify what we mean by writing quality. It is important to note, however, that e-rater models are optimized to predict human holistic ratings, and thus the component features selected for use in e-rater were selected to maximize prediction, not necessarily to fully represent all aspects of writing cognition. It may therefore be necessary to supplement the features used in an AES system with additional features to support a detailed cognitive analysis.

## 4. Predicting the Developmental Level Of Student Writing

Another advantage of AES is the possibility of applying a single consistent metric. In many large-scale assessments, score comparability is critical, e.g., from Time 1 to Time 2, from Form 1 to Form 2, from Grade 6 to Grade 8. This is particularly challenging for writing assessments, since human scoring techniques require comparable judgments from different raters reading different essays at different times under varying conditions. Since AES techniques are consistent by definition, being mechanically computed from the written text features, it is likely that automatically assigned scores can be used to calibrate assessments across grades and other population differences. Of course, this possibility depends upon consistent developmental trends in the features underlying AES, which is entirely an empirical question.

Attali and Powers (2008) addressed this question by recruiting a national sample of students in classes at grades 4, 6, 8, 10, and 12. These classes were randomly assigned to a condition, which varied the writing mode (persuasive or descriptive) and grade level of the prompt. Students wrote two essays in the assigned mode, responding to one prompt from their current grade level and another from an adjacent grade level. For example, a 10$^{th}$ grade student might respond to 10$^{th}$ and 8$^{th}$ grade-level prompts. Essays were collected from more than 12,000 students; the usable sample contained 34,630

essays. Thus, the authors collected a very large corpus of student writing, from a cross-section of the U.S. student population, from primary and secondary students, composing multiple texts. In the past, this large representative corpus might have yielded few insights into student writing, due to the inherent limitations of human scoring; however, AES makes possible automated analyses of even large corpora, for relatively fine-grained linguistic features.

Essays were scored with e-rater version 2.0, using a scoring model, in which the 8 e-rater features were weighted equally. Predictable patterns of performance were observed, with students progressively increasing their performance on the underlying features across grades, and were thus able to develop grade-level norms that located students on a scale of writing development, at least under the writing conditions examined (30 minute tasks focused on very general prompts).

Attali and Powers' (2008) analyses indicate that e-rater scores have developmental validity as predictors of writing quality. The relationship between e-rater score and writing skill may be indirect, since many of the predictors of writing quality may represent prerequisite skills. Nonetheless, feature norms can be established by examining AES features at different grade levels in a large corpus of student writing, and such patterns provide fundamental data for any theory of writing quality and writing skill.

## 5. Identifying Dimensions of Linguistic Variation in Student Essays

As the preceding sections suggest, automated linguistic analysis represents a promising methodology for corpus analysis. A wide variety of features can be collected from a corpus of student writing, and their distributions can then be examined across grade levels or other dimensions of interest. Factor analysis can then be used to identify major dimensions of covariance. This kind of analysis is comparable in method to techniques used by Biber and his colleagues (1988; 1995a; 2004). Since all texts are written by somebody, every text corpus, by its patterns of variation, reveals something about the choices writers make. Sheehan, Kostin, and Futagi (2007) examined a corpus of source documents for potential use by ETS testing programs, where the intent was to filter texts so that they correctly instantiated genres of interest.[4] Six factors were extracted: spoken language, academic discourse, overt expression of persuasion, oppositional reasoning, sentence complexity, and unfamiliar vocabulary. Deane, Sheehan, Sabatini, Futagi, and Kostin (2006) examined variations in a set of 3rd through 6th grade texts drawn from a large corpus of materials typically used in school. Nine factors were extracted: spoken language, oppositional reasoning, academic discourse, causal reasoning, overt expression of persuasion, sentence complexity, word familiarity, impersonal reference and numeric vocabulary.[5] Sheehan, Kostin, Futagi, and Sabatini (2007) examined a corpus of texts intended for readers ranging from early primary to high school and developed a similar factor analysis that yielded nine factors which included academic style, sentence complexity, vocabulary difficulty, subordination and oppositional

reasoning. These five factors were used to develop predictive grade-level models of readability by genre in which academic style, subordination and oppositional reasoning contributed to the prediction of readability alongside the sentence complexity, and vocabulary difficulty factors more typically used in readability measures.

The same kind of analysis can be applied to the features used in AES. While validating their developmental scale, Attali and Powers (2008) also examined the internal structure of e-rater scores on student essays. As described above, the authors collected a very large, representative sample of student writing. They conducted exploratory and confirmatory factor analyses using e-rater variables, e.g., Grammar, Usage, Mechanics, Style, Vocabulary, and Word Length (but replacing e-rater's Organization and Development features with essay length).The authors found that the better fitting model depended upon student grade. A two-factor model provided a better fit for grades 4 and 6, while a three-factor model better fit grades 8, 10, and 12. The authors interpreted the three-factor model as: fluency (essay length, Style), sentence conventions (Grammar, Usage, & Mechanics), and word choice (Vocabulary & Word Length). The two-factor model represented a merging of fluency and sentence conventions.

To be most meaningful, linguistic dimensions should be stable. In the case of Attali and Powers' (2008) results, we would want evidence that the linguistic dimensions did not reflect some particularity of the analysis, such as how e-rater aggregates microfeatures into features or the selection of one feature (such as vocabulary frequency) over another (such as abstractness and imageability). Accordingly, we replicated Attali and Powers (2008) exploratory and confirmatory factor analyses on a subset of that corpus containing 17,586 student essays[6]. Instead of e-rater's feature scores (used by Attali & Powers), we used a large array of linguistic measures, hoping by casting a wider net to obtain a more nuanced picture of writing development.

First, we used individual e-rater microfeatures (see Figure 1). Second, we incorporated a number of features derived from ETS' Sourcefinder (Sheehan et al., 2006); a tool developed for selecting grade appropriate reading passages for inclusion in ETS tests. After a review of the data, some features that did not perform well were eliminated. This process left 40 linguistic features, which formed the basis of a more detailed developmental study of automatically-collected features associated with writing quality.

### 5.1 Data Preparation

Our selection from the Attali and Powers' (2008) dataset was divided into four subsets, based upon essay order. During data collection, the authors counter-balanced the order in which students encountered the two main conditions (i.e., genre and grade-level of prompt). For our purposes, these subsets provide a convenient means for cross-validating our results. Since some participants were lost as the Attali-Powers study proceeded, the first essay order contained 5,150 essays after outliers were eliminated, the second, 4,940 essays; the third, 4,162 essays, and the fourth, 3,284. Each essay

order contained either a persuasive essay or a descriptive essay. The essays were processed using ETS' natural language processing software to identify candidate features. Some of the features thus selected were too sparse to reliably be included in a factor analysis, and were excluded; in particular, the analysis excluded any feature with nonzero values in less than 5% of cases. Where possible we preferentially included features that positively correlated with grade level and human essay quality ratings.

## 5.2 Exploratory Factor Analysis

Exploratory factor analysis (EFA i.e., principle components-based factor analysis with Promax rotation) was performed upon each subset of essays. The results of each EFA was used to cross-check the others, and thus to obtain the most interpretable set of factors. A factor structure with ten factors replicated across all four essay orders (see Table 1).

**Table 1. Exploratory Factor Analyses of the Attali/Powers Dataset**[7]

| Feature | Subset 1 | Subset 2 | Subset 3 | Subset 4 |
|---|---|---|---|---|
| **Dimension 1: Academic Orientation** | | | | |
| Avg. no. syllables in a word | +.93 | +.96 | +.92 | +.93 |
| Nominalizations (*-tion, -ment, -ness, -ity)* | +.81 | +.80 | +.78 | +.76 |
| Academic Verbs (*apply, develop, indicate* etc) | +.78 | +.73 | +.79 | +.75 |
| Academic Words (Coxhead) | +.76 | +.74 | +.79 | +.75 |
| Abstract Nouns (*existence, progress* etc) | +.69 | +.57 | +.66 | +.63 |
| Passive Verbs | +.52 | +.52 | +.46 | +.43 |
| Median Word Frequency | -.63 | -.60 | -.60 | -.58 |
| Dale List of Common Words | -.84 | -.85 | -.82 | -.86 |
| Imageability Score [MRC database] | -.85 | -.85 | -.81 | -.88 |
| **Dimension 2: Noun-Centered Text** | | | | |
| Definite determiners (log per 1000 words) | +.92 | +.89 | +.98 | +.87 |
| Wrong, missing or extraneous articles | +.68 | +.72 | +.70 | +.74 |
| Noun/Verb Ratio | +.56 | +.60 | +.60 | +.56 |
| Nouns (log per 1000 words) | +.52 | +.56 | +.55 | +.53 |
| Document length in words | +.35 | +.38 | +.34 | +.38 |
| **Dimension 3: Sentence Complexity** | | | | |
| Verbs (log per 1000 words) | +.96 | +.92 | +.98 | +.94 |
| Average sentence length in words | +.93 | +.89 | +.95 | +.92 |
| Prepositions (log per 1000 words) | +.48 | +.49 | +.48 | +.58 |
| **Dimension 4: Spoken Style** | | | | |
| Mental State Verbs (appreciate, care, feel etc) | +.78 | +.84 | +.78 | +.79 |
| Conversation Verbs (get, know, put etc.) | +.67 | +.68 | +.74 | +.74 |
| First Person Singular Pronouns | +.30 | +.37 | +.41 | +.45 |
| Noun/Verb Ratio | -.36 | -.33 | -.33 | -.35 |
| Attributive Adjectives (log per 1000 words) | -.75 | -.58 | -.73 | -.75 |
| **Dimension 5: Overt Expression of Persuasion** | | | | |
| Predictive Modals (will, would etc.) | +.84 | +.86 | +.87 | +.86 |
| Conditional Subordinators (if, unless etc) | +.74 | +.75 | +.80 | +.71 |

| | | | | |
|---|---|---|---|---|
| Present Tense | -.46 | -.61 | -.54 | -.66 |
| **Dimension 6: Elaboration** | | | | |
| Document length in words | +.67 | +.66 | +.66 | +.56 |
| Indefinite pronouns (someone, anyone, etc.) | +.62 | +.42 | +.71 | +.53 |
| Adversative Conjunctions (alternatively, etc.) | +.59 | +.66 | +.49 | +.71 |
| Concessive Subordinators (although, though) | +.40 | +.45 | +.37 | +.63 |
| **Dimension 7: Narrative Style** | | | | |
| Past Tense Verbs | +.78 | +.79 | +.78 | +.83 |
| Past Perfect Aspect Verbs | +.72 | +.65 | +.67 | +.64 |
| Third Person Singular Pronouns | +.33 | +.43 | +.40 | +.54 |
| Pesent Tense Verbs | -.52 | -.43 | -.55 | -.40 |
| **Dimension 8: Orthographic Errors** | | | | |
| Contraction/apostrophe errors | 0.71 | +.70 | +.72 | +.70 |
| Didn't capitalize proper noun | 0.64 | +.65 | +.74 | +.62 |
| Spelling | 0.62 | +.63 | +.53 | +.64 |
| Confusion of Homophones | 0.44 | +.39 | +.21 | +.39 |
| **Dimension 9: Verb Errors** | | | | |
| Ill-formed Verb | 0.68 | +.67 | +.52 | +.82 |
| Subject/Verb Agreement | 0.64 | +.67 | +.63 | +.55 |
| Proofread This | 0.51 | +.45 | +.67 | +.44 |
| **Dimension 10: Comma Errors** | | | | |
| Comma Errors | +.93 | +.92 | +.89 | +.91 |

Many of the dimensions identified in the present EFAs closely resemble dimensions identified in previous studies. In particular Academic Orientation, Sentence Complexity, Spoken Style, Overt Expression of Persuasion, and Narrative Style are very similar to factors uncovered in the genre analyses performed by Biber and his colleagues (Biber, 1986; Biber, 1988; Biber et al., 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Reppen, 2001). This resemblance suggest that much of the variation in student's writing reflects their growing ability to produce texts that approximate the structure of well-formed, academic, written language.

Three of the identified factors, i.e., elaboration, orthographic errors, and verb errors, do not appear in any of Biber's analyses. Differences between Biber's purpose and the purpose of the present investigation provide a straight-forward explanation. In the elaboration factor, essay length (number of words) accounts for the most variance. In contrast, Biber controls for variation in text length by selecting the same number of words from each source text. Further, Biber's analyses could not have revealed factors for orthographic errors or verb errors, since his analyses focused on edited text and therefore did not include any measure for errors. In the present results, these two factors mostly comprise microfeatures aggregated into one of two e-rater features, mechanics and grammar.

To what extent are these linguistic dimensions meaningful for understanding students' writing skills? One way to address this question is to examine the relationship to essay quality and human scoring, discussed at length above. In holistic scoring,

human raters form an "overall impression" guided by a rubric that specifies multiple dimensions of essay quality. Another scoring method, analytic trait scoring makes these quality dimensions explicit. Both require careful training to avoid issues of inter-rater reliability

Here, AES may have some advantage over human scoring. In the context of scoring student essays, human reading comprehension appears well-adapted to gist understanding, which may serve holistic scoring well, but may not efficiently support the fine-grained analyses required in analytic trait scoring. While automated scoring techniques have the potential for carrying out fine-grained analysis, they can only do so usefully to the extent that they measure the kinds of quality traits identified by humans.

To what extent do the results of the EFA, listed above, map to recognizable traits of essay quality? To evaluate this question, we examined how well the factor scores correlated with grade level and human holistic scores (see Table 2).[8] For most of the factors (Academic Orientation, Noun-Centered Text, Spoken Style, Overt Expression of Persuasion, Narrative Style, Orthographic Errors, & Verb Errors), the relationship to Grade Level and Human Scores appears roughly comparable, in terms of correlation strength. These factors related strongly to both Grade Level and Human Quality Scores. This pattern of relationship suggests a general developmental progression, which may indicate maturation and continued acquisition of verbal skills and/or literacy skills.

In contrast, for two dimensions, Sentence Complexity and Elaboration, there is a large discrepancy in the pattern of associations between grade level scores and essay quality. Sentence Complexity correlated much more strongly with grade level than with essay quality. Writing manuals suggest that writers use a mix of simple and complex sentences (Strunk, 2000); thus we might expect a weak relationship between Sentence Complexity and Human Score, since a good writer might be able to produce complex sentences, but choose a simpler sentence structure for the sake of clarity.

On the other hand, Elaboration correlated much more strongly with human scores than with grade level. On this dimension, essay length accounts for most of the variance, perhaps because the ability to produce extended text, such as an essay, presupposes a certain level of fluency. While the length of an essay may be a rather ambiguous proxy of essay quality, since writing manuals excoriate verbosity and praise concision (cf., Strunk, 2000), longer essays are consistently assessed more positively both by humans and by AES systems. Some writing researchers argue that this relationship reflects the relative fluency of text production skills (Chenoweth & Hayes, 2001; Ransdell & Levy, 1996).

Besides essay length, the Elaboration dimension includes other measures (indefinite pronouns, adversative conjunctions, and concessive subordinators) that signal abstraction, negation, and subordination. Such features typically appear in complexly structured texts. Thus, the Elaboration dimension seems to capture more than simple fluency, and may also partially reflect the ability to produce complexly structured texts.

**Table 2. Correlations of Each Factor With Grade Level and Human Scores Where Available. Almost all correlations are significant at the 0.01 level; all are significant at the 0.05 level[9]**

| Factor | U.S. Grade Level (4th, 6th, 8th, 10th, or 12th grades) (N=17,586) | Human Scores (Persuasive Essay Prompt) (N=589) | Human Scores (Expository Essay Prompt) (N=568) |
|---|---|---|---|
| Academic Orientation | .52 | .47 | .49 |
| Noun-Centered Text | .30 | .42 | .27 |
| Sentence Complexity | .35 | .12 | .17 |
| Spoken Style | -.24 | -.29 | .03 |
| Overt Expression of Persuasion | -.14 | .10 | .05 |
| Elaboration | .29 | .48 | .53 |
| Narrative Style | .11 | .21 | .24 |
| Orthographic Errors | -.19 | -.24 | -.28 |
| Verb Errors | -.08 | -.09 | -.13 |

The general picture is that with age, instruction, and practice, student essays become more overtly academic and less oral in vocabulary and style, with a higher concentration of nominal text and a greater degree of elaboration, but with fewer orthographic or grammatical errors.[10] Judgments of essay quality consistently are dominated by measures of elaboration and academic language; with descriptive essays allowing more oral style features than typically appear in persuasive essays. These results are suggestive, but the correlation between factors and judgments of essay quality should be viewed as preliminary and exploratory, suggesting hypotheses to be tested. In particular, there is a clear need to examine a larger human-scored corpus to elucidate interactions among essay quality, general verbal development, and factors based upon NLP features

## 5.3 Second Order Factor Analysis

The factor analysis yielded ten factors, which could reflect similarities or overlaps in the information carried by nominally different features. Therefore, a second order factor analysis was performed on seven of the ten factors identified in the EFA. The two genre factors were excluded (narrative style, overt expression of persuasion), because they had no clear correlates in the Attali-Powers factor analysis, and appeared to have very little association with measures of essay quality. The tenth factor, which only attracted a single feature, was also excluded. Table 3 shows the pattern matrices that resulted over

all four essay orders. Our purpose was to correct for the disaggregation we had performed by examining e-rater microfeatures directly. There was significant covariance among the ten factors obtained in the first-order analysis, and we suspected that further analysis would reveal a pattern rather similar to that obtained by Attali and Powers (2008).

**Table 3. Second Order factor Analyses Over the 4 Essay Orders on 7 of the 10 Factors Identified by the Exploratory Factor Analysis**

| Factor | Component 1 by essay order | | | | Component 2 by essay order | | | | Component 3 by essay order | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Academic Orientation | **.80** | **.82** | **.83** | **.84** | .17 | .15 | .13 | .17 | -.17 | .15 | .14 | .10 |
| Noun-Centered Text | **.86** | **.85** | **.85** | **.84** | .08 | .04 | .10 | -.09 | -.07 | .04 | .12 | .21 |
| Sentence Complexity | .30 | .21 | .23 | .32 | **.76** | **.79** | **.81** | **.56** | -.01 | .79 | .05 | .34 |
| Spoken Style | **-.80** | **-.79** | **-.72** | **-.72** | .17 | .22 | .34 | .13 | -.11 | -.05 | -.01 | .23 |
| Elaboration | **.80** | **.79** | **.72** | **.72** | .17 | .22 | .34 | .13 | .03 | .05 | .01 | .23 |
| Orthographic Accuracy | -.21 | -.34 | -.26 | -.32 | **.74** | **.68** | **.72** | **.67** | **.79** | .02 | -.02 | .10 |
| Verb Errors | -.14 | -.07 | -.06 | -.02 | -.26 | -.36 | -.32 | **-.85** | **.74** | **.61** | **.69** | .27 |
| | -.12 | -.06 | -.04 | -.06 | -.33 | -.26 | -.25 | -.10 | | **.87** | **.86** | **.93** |

The second order analysis yielded a three-factor solution that was exactly the same over three of the four essay orders, and differed only in one assignment in the fourth order. In the dominant pattern, the first factor (academic orientation, noun-centered text, spoken style) is readily interpretable as a combination of vocabulary usage and syntactic style[11], spanning the range between prototypically spoken, oral language and prototypically academic, written language. The second factor (sentence complexity, elaboration) is readily interpretable as a fluency dimension, though it could also be interpreted as the ability to produce complex, well-structured texts. And the third factor (orthographic accuracy, verb errors) is readily interpretable as an accuracy dimension.

These three factors correspond roughly to the three factors identified by Attali and Powers (2008), with the fluency factor in that analysis corresponding to component 2 in Table 3, the conventions factor, to component 3, and the word choice factor to the academic orientation factor. In particular, these tables indicate obvious associations between e-rater features and the three macro-factors that we have identified. Thus, the e-rater vocabulary features (median word frequency and average word length) correlate most strongly with the Academic Orientation, Noun-Centered Text, and Spoken Style factors, i.e., with the spoken vs. academic 2nd order factor. As might be expected given the sharing of features, the orthographic accuracy factor correlates most strongly with

the mechanics feature, while the verb error factor correlates most strongly with the grammar feature.

While orthographic accuracy patterned differently in one subset where it aligned with factor two instead of factor three[12], the overall trend is very clear. The Attali-Powers developmental data clearly reflects three components even when a wide range of features are included. These factors appear to correspond roughly to the following abilities:

1. The ability to produce documents fluently, with appropriately complex sentences and evidence of appropriate elaboration of text structure
2. The ability to adopt vocabulary to an appropriately academic style, with more typically written grammatical patterns and an avoidance of typically oral patterns.
3. The ability to maintain conventional patterns of grammar, mechanics and usage.

### 5.4 Analysis of excluded features

Among the features available for our analysis were a number of grammatical categories, such as attributive adjectives and function words. In our data, many of these features were present in less than 5% of cases; for this reason, they were excluded from the EFA. Many of the remaining features of this type had relatively small (though significant) weights in the factor analysis; for the sake of clarity, these were not presented in the results (Table 1). Yet many of them also exhibited highly significant correlations with grade level and essay score. It seemed possible that they might provide a measure of one aspect of writing quality – sentence variety—if they were aggregated into a single feature. To examine this possibility, we conducted multiple regression analyses over the training data, using these excluded features to predict grade level after document length was factored out[13]. In the results, a number of these features strongly weighted positively (w = .27 to .64) in the regression equation, such as wh-determiners (*whose, which*), academic downtoners (*barely, hardly*, etc.), perfect aspect verb forms, focus adverbs (*only, even*), negative universal quantifiers (*never, no one*, etc.), and wh-adverbs (*where, when, why*). Other features weighted negatively (w = -.21 to -.39), including emotion words, mental state verbs, 2nd person pronouns, 3rd person pronouns, and communication verbs.

These results are consistent with the general picture obtained above: the model assigns positive weights to a variety of syntactic constructions associated with academic discourse, while assigning negative weights to pronouns and other cues indicating a relatively oral style. These results suggest a developmental trajectory, in which students gradually master the syntactic and lexical choices characteristic of academic discourse.

Thus, using NLP techniques to analyze student essay corpora suggests intriguing hypotheses about students' writing skills. Many features reflect general verbal maturity, as suggested by their correlations with grade level. Our results suggest that one important dimension of this development reflects an increasing ability to adopt an academic register in writing.

Most of the identified dimensions had roughly comparable correlations both with grade level and with human holistic judgments of essay quality, suggesting that they represent skills that develop along an expected trajectory, wherein more mature students better approximate the lexical and stylistic characteristics of higher-quality texts. However this expected trajectory was not found for two dimensions, sentence complexity and elaboration. In the first EFA, elaboration correlated strongly with human scores, but more weakly with grade level; while sentence complexity showed the reverse pattern. Yet, in the second order FA, these two factors positively loaded on the same component, suggesting that they have something in common. Since this component broadly encompasses the ability to compose extended text with more complex sentences, it potentially makes a very important contribution to overall writing skill. More research is needed to better understand the development of this component.

The data considered in our analyses represents a very specific sample of student writing: an impromptu essay composed under timed conditions. In order to generalize our findings, it will be necessary to replicate these findings on a broader range of writing tasks, in different genres, under a variety of cognitive conditions, across time. But this is where the automated techniques of corpus analysis come into their own: they are limited only by the availability of appropriate corpora, since they can be scaled up to as large a corpus as necessary.

## 6. Going Beyond NLP Analysis of Writing Quality

If a student's essay receives a high quality rating, we may infer that he or she has sufficient skills to write an essay. However, what can we infer from low ratings? Perhaps the student had little opportunity to plan or revise. Perhaps the student struggled with typing or spelling. Under some conditions even an expert writer may produce a low quality essay. Ultimately, the final essay may reveal little about the processes that went into making it. Therefore a corpus consisting only of final drafts may not yield all the insights we might hope for. To understand strengths and weaknesses of student writing, we may need to look beyond the final essay in various ways, to consider both the process and the product. For instance, we may examine choices the writer makes during the writing process; we may compare automatically-scored features of multiple drafts; we may consider data from keystroke logs, and a variety of other techniques. However, each of these techniques can be enriched by combining it with automated NLP analysis. While our research to date has focused upon predicting human judgments of essay quality, the same techniques, combined with process data, can be used to enrich and disambiguate a cognitive account of writing skill.

### 6.1 Evaluating students' engagement with the writing process

Students' written texts are the product of a sequence of automatic and strategic cognitive processes. A student may begin writing with no plan in mind, letting one idea

prompt the next; or he may adopt a more complex strategy, such as stopping to first to prepare an outline. Students may write through from beginning to end, and stop there, or they may engage in cyclical revision. The advent of online literacy environments makes it relatively easy to collect information about these behaviors. In 2000, Educational Testing Service introduced the Criterion® Online Writing Evaluation system, a writing tool designed to support writing instruction. Initially, Criterion's key feature was providing students with Internet-based scoring and comparison with benchmark student essays, but later was adapted to provide feedback on grammar, usage, mechanics, style, organization and development. In Criterion, students can compose essays in response to a range of essay topics, including teacher specified topics. After submitting the draft, Criterion identifies possible problems or errors in the essay draft and (optionally) a holistic score. Writers may then proceed to revise and edit. Thus, Criterion can be seen as a tool for promoting writing practice. Currently it is used by hundreds of thousands of students throughout the world.

From the Criterion database we extracted usage data on a large sample of essays (n = 185,964) composed by U.S. students, grades 6 through 12. If the student chose to prewrite and/or revise, then information on these activities was included in the data. Data included measures of prewriting (time on-task, number of planning words), initial draft (time on-task, number of words, number of errors, predicted holistic score), and final draft (time on-task, change in number of words, change in number of errors, change in holistic score).[14]

About one half of the time (51%), students produced a single draft without prewriting or revision. In very few cases (2.5%), did students plan, draft, and revise. Among projects that included revision, we conducted two analyses. First, we examined the change from first to last draft. We regressed variables of change (i.e., # of words added, # of errors corrected, and time spend revising) onto a measure of overall improvement (i.e., difference in holistic score from final to first draft). The strongest predictor of change in holistic score was word count ($ß = .12$; $t = 65.62$), suggesting that students improved most by producing more fully-elaborated texts. Second, we examined planning. When students used the planning tools supplied with Criterion, they tended to spend more time on-task and produce higher quality essays (as measured by e-rater score). These results are consistent with studies indicating a similar positive effect for advance planning (e.g., Kellogg, 1987; Quinlan, 2004). It is thus possible to collect corpora that include information about prewriting, drafting, and revising, and it thus may be possible to obtain additional insights into students' writing processes.

Currently our Criterion data is primarily 'found' data (byproducts of operational use.) In future work, we intend to conduct well-controlled studies involving systematic manipulations of these elements of the writing process. We are particularly interested in whether systematic effects upon the final quality of essays can be induced by various interventions, such as training students in the use of planning tools. Given the availability of automated assessments of essay quality, with known, strong correlations

with human judgments, AES can be used to support larger-scale studies of the effects of interventions, or other manipulations of students' habitual writing processes, on essay quality.

## 6.2 Measuring Text-production

Automated corpus analysis may also support fine-grained studies of text production. At the most basic level, writing skill presupposes the ability to translate ideas into words, arrange those words grammatically, and then transcribe them. Hayes and Flower (1980) called this process "translating". When students compose on a personal computer, information can be captured about every keystroke and analyzed to provide insights into students' basic text production skills. Will, Nottbusch, and Weingarten (2006) found skilled writers had significantly longer latencies between keystrokes for low-frequency than for high-frequency words. Keystroke latencies were also prolonged at syllable and morpheme boundaries. The authors conclude that sub-word-level processes are involved in text production, even in highly fluent text production (such as by college students).

Competent writing depends upon reasonable fluency of text production, and poor performance may be caused by dysfluent text production. For example, students with learning disabilities (LD) typically compose shorter texts than normally achieving students (Nodine, Barenbaum, & Newcomer, 1985), yet produce more spelling, capitalization and punctuation errors (Moran, 1981; Poteet, 1979). Students with LD also fail to monitor their texts for inconsistencies unless prompted (Bos, Anders, Swanson, & Keogh, 1990) and demonstrate an underdeveloped sense of when a composition is complete (Englert, Hiebert, & Stewart, 1988).

Here again, the availability of automatic essay scoring makes it possible to envisage much larger-scale studies than would otherwise be possible. Given features representing dynamic writing behaviors, and an AES system, it becomes possible to explore how profiles of writing behavior correlate with essay quality, over very large numbers of students. Conversely, it is possible to manipulate writing conditions, observe the effect of manipulations on behavioral variables, and examine how the manipulations in turn affect essay quality as measured by an automated essay score. When combined with keystroke logging programs, it will be possible to examine the interaction of essay quality with students' text production skills unobtrusively. Fluency and accuracy of text production represent important measures of students' abilities to get words on the page, and thus can provide indicators of progress toward competent writing.

While we have not yet conducted studies of this type, we have preliminary data that suggest it can be fruitful to combine NLP feature analysis with measures of text production. As part of pilot testing of a writing assessment being developed at Educational Testing Service, we obtained a small number of student responses for which we were able to collect both keystroke logs and a full complement of automatically-calculated NLP features for 40 student essays. These essays were part of a

pilot administration of a new essay test in which students were asked to write a persuasive writing prompt after being presented with preliminary inquiry and analysis questions focused on related reading material. Students were drawn from three middle schools at a school district in a northeastern state, covering a mix of urban, suburban and rural students, including a significant number of English language learners drawn primarily from a refugee population. About 120 students were tested in total, but keystroke logging was implemented only for a subset, primarily because of operational conditions that required the keystroke logging facility to be deactivated partway through the administration. Thus the dataset we obtained should be viewed only as a preliminary sample. We are currently collecting keystroke logs as part of a large national data collection for the same test development effort (four samples with more than 1,000 students per essay), and we intend in future work to examine correlations among behavioral features, NLP features, and measurements of essay quality. But as an initial exploration we examined the same correlations in this small preliminary sample.

We conducted multiple regression analyses to determine which features best predicted human ratings of writing quality, including a variety of timing-based features (burst length, mean pause between characters, mean pause between words, mean pause between sentences, average length of time spent backspacing.) The best model obtained (allowing all features to compete with one another in a stepwise regression) is that shown in Table 4, which slightly outperforms a standard e-rater model (for which the Adjusted R-Square is .83): the overall measures (R, R-Square, and Adjusted R-Square) reflect a high degree of accuracy in predicting human ratings. The standard coefficients (measuring how much each feature contributes to the overall prediction of essay score) reflect a fairly even division among three features: organization, pause length, and mechanics. Higher organization and mechanics scores predict higher essay qualities, as do relatively short pauses between words. Since this is a stepwise regression, other features have been excluded because they are redundant (that is, they largely covary with the selected features)

**Table 4. A Stepwise Regression Model Predicting Essay Quality (holistic human scores) From a Combination of NLP and Behavioral Features**

| ***R*** | ***R*²** | **Adjusted *R*²** |
|---|---|---|
| **.93** | **.88** | **.86** |

| **Feature** | **Standard Coefficients (Beta)** | **Significance** |
|---|---|---|
| E-rater Organization Feature | .538 | <.001 |
| Mean pause between words | -.436 | <.001 |
| E-Rater Mechanics Feature | .311 | <.01 |

The largest difference between this model and a standard e-rater scoring model is that one text production measure (mean pause between words) has supplanted the development feature (average length of discourse units), perhaps because both measure

an underlying dimension: fluency of text production. While the model is based on far too small a sample to draw any strong conclusions, it does suggest that useful information could be extracted from large text corpora by combining automated scoring with the extraction of behavioral features. We expect to explore this question in detail in future work. We are particularly interested in the extent to which differences in patterns of timing might give us access to information about the efficiency of text production and the amount of time writers are devoting to strategic planning.

## 7. Conclusion

Corpus analysis can reveal much about the development of students' writing skills. For several years, research at Educational Testing Service has focused on automatic analysis of student writing, the results of which demonstrate the feasibility of a) predicting human judgments of essay quality, b) placing essays on a developmental scale, and c) identifying the linguistic dimensions underlying student writing. While these analyses have been successful, leading to operational use of AES in writing assessment, we suspect that there are limits to the amount of information that can be wrung from a student essay without additional sources of data. In the quest to understand students' writing skills, we are beginning to explore corpus analyses that include both the writing process and the written product.

By broadening our definition of corpora, we are capturing some of the dynamics of written composition. Online writing environments, combined with NLP automated scoring technology, means that we can predict human scores for very large collections of student writing, and can combine that with information about the time course of text production and with detailed profiles of students' prewriting, drafting, and revising. At this point, our research into capturing the dynamics of composing has progressed only far enough to present intriguing possibilities. There is clearly considerable scope for research into the nature of writing that takes advantage of NLP techniques, and we intend to do so in ongoing work in support of essay scoring. However, the same techniques have obvious applications for researchers interested in the nature of writing quality, and its connection to underlying cognitive processes.

## References

Attali, Y., & Burstein, J. (2006a). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment, 4*(3). Available from http://www.jtla.org.

Attali, Y., & Burstein, J. (2006b). Automated essay scoring with E-rater V. 2.0. *The Journal of Technology, Learning, and Assessment, 4*(3), 13-18.

Attali, Y., & Powers, D. (2008). *A developmental writing scale*. Princeton, NJ: Educational Testing Service.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language, 62*(2), 384-414. doi: 10.2307/414678

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511621024

Biber, D. (1995a). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511519871

Biber, D. (1995b). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511804489

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., et al. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus* (ETS TOEFL Monograph Series No. MS-25). Princeton, NJ: Educational Testing Service.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman grammar of spoken and written English. Harlow, Essex: Pearson.

Bos, C. S., Anders, P. L., Swanson, H. L., & Keogh, B. K. (1990). Toward an interactive model: Teaching text-based concepts to learning disabled students. In B. Keogh & H. L. Swanson (Eds.), *Learning disabilities: Theoretical and research issues* (pp. 247-261). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill* (No. 0-87447-280-6). New York, NY: College Entrance Examination Board.

Burstein, J., Chodorow, M., & Leacock, C. (2003, August). *Criterion℠ Online Essay Evaluation: An application for automated evaluation of student essays.* Paper presented at the Fifteenth Annual Conference on IAAAI'03, Acapulco, Mexico.

Burstein, J., & Higgins, D. (2005, July). *Advanced capabilities for evaluating student writing: Detecting off-topic essays without topic-specific training*. Paper presented at the International Conference on Artificial Intelligence in Education, Amsterdam, Netherlands.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). *Computer analysis of essays*. Paper presented at the Proceedings of the NCME Symposium on Automated Scoring, San Diego, CA.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., et al. (1998). Automated scoring using a hybrid feature identification technique. In C. Boitet & P. Whitelock (Eds.), *Proceedings of the Annual Meeting of the Association of Computational Linguistics: Vol. 1* (pp. 206-210). Morristown, NJ: Association for Computational Linguistics.

Burstein, J., & Marcu, D. (2000). Benefits of modularity in an automated essay scoring system. In *Proceedings of the workshop on using toolsets and architectures to build NLP systems, 18th international conference on computational linguistics*. Luxembourg: Association for Computation Linguistics.

Burstein, J., & Marcu, D. (2003). A machine learning approach for identification thesis and conclusion statements in student essays. *Computers & the Humanities, 37*(4), 455-467. doi: 10.1023/A:1025746505971

Burstein, J., Marcu, D., Andreyev, S., & Chodorow, M. (2001, July). *Towards automatic classification of discourse elements in essays*. Paper presented at the Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France.

Chenoweth, N., & Hayes, J. R. (2001). Fluency in writing. *Written Communication, 18*(1), 80-98. doi: 10.1177/0741088301018001004

Chodorow, M., & Leacock, C. (2000, April 29-May 4). *An unsupervised method for detecting grammatical errors*. Paper presented at the Proceedings of the First Conference on North American chapter of the Association for Computational Linguistics, Seattle, WA.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanovich.

Crowhurst, M. (1980). Syntactic complexity and teachers' quality ratings of narrations and arguments. *Research in the Teaching of English, 14*(3), 223.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Usman, E., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL*. Princeton, NJ: Educational Testing Service.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin, 27*(1), 11-28.

Dale, E., & Tyler, R. W. (1934). A study of the factors influencing the difficult of reading materials for adults of limited reading ability. *The Library Quarterly, 4*(3), 384-412. doi: 10.1086/613490

Deane, P., Sheehan, K., Sabatini, J., Futagi, Y., & Kostin, I. (2006). Differences in text structure and its implications for assessment of struggling readers. *Scientific Studies of Reading, 10*(3), 257-275. doi: 10.1207/s1532799xssr1003_4

Diederich, P. B. (1974). *Measuring growth in English*. Urbana, IL: National Council of Teachers of English.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in the judgement of writing quality* (ETS Research Bulletin No. RB-61-15). Princeton, NJ: Educational Testing Service.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. New York: Peter Lang.

Englert, C. S., Hiebert, E. H., & Stewart, S. R. (1988). Detecting and correcting inconsistencies in the monitoring of expository prose. *The Journal of Educational Research, 81*(4), 221-227.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233. doi: 10.1037/h0057532

Flower, L., & Hayes, J. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365-387. doi: 10.2307/356600

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*(2), 285-307. doi: 10.1080/01638539809545029

Foltz, P., Laham, D., & Landauer, T. K. (1999a). The intelligent essay assessor: Applications to educational technology [Electronic Version]. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*. Retrieved June 30, 2003 from http://imej.wfu.edu/articles/1999/2/04/printver.asp.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999b). Automated essay scoring: Applications to educational technology. In B. Collis & R. Oliver (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 1999* (pp. 939-944). Chesapeake, VA: AACE.

French, J. W. (1961). *Schools of thought in judging excellence of English themes*. Reprint from the Proceedings of the Invitational Conference on Testing Procedures, Princeton, NJ: Educational Testing Service, 1962.

Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review, 35*(3), 435-450.

Godshalk, F. I., College Entrance Examination Board, & et al. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Hayes, J. R., & Flower. (1980). Identifying the organization of writing processes. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum Associates.

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004, May). *Evaluating multiple aspects of coherence in student essays*. Paper presented at the Proceedings of the 2004 HLT/NAACL, Boston, MA.

Hunt, K. W. (1970). Syntactic maturity in schoolchildren and adults. *Monographs of the Society for Research in Child Development, 35*(1), pp. iii-iv; 1-67. doi: 10.2307/1165818

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In B. Huot & M. Williamson (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). Cresskill, NJ: Hampton press.

Huot, B. (1996). Computers and assessment: Understanding two technologies. *Computers and Composition, 13*(2), 231-243. doi: 10.1016/S8755-4615(96)90012-2

Kellogg, R. T. (1987). Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory & Cognition, 15*(3), 256-266. doi: 10.3758/BF03197724

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57-71). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

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*(2), 211-240. doi: 10.1037/0033-295X.104.2.211

Lee, Y. W., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL CBT essays: Scores from humans and e-rater*. Princeton, NJ: Educational Testing Service.

Lively, B. A., & Pressey, S. L. (1923). A method for measuring the "vocabulary burden" of textbooks. *Educational Administration and Supervision, 9*, 389-398.

Loban, W. (1976). *Language development: Kindergarten through grade twelve* (No. NCTE Committee on Research Report No. 18). Urbana, IL: National Council of Teachers of English.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276. doi: 10.1191/0265532202lt230oa

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8*(3), 299-325. doi: 10.1007/BF01464076

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-76. doi: 10.1177/026553229000700105

Moran, M. R. (1981). *A comparison of formal features of written language of learning disabled, low-achieving and achieving secondary students*. Lawrence,KN: Kansas University.

Nodine, B. F., Barenbaum, E., & Newcomer, P. (1985). Story composition by learning disabled, reading disabled, and normal children. *Learning Disability Quarterly, 8*(3), 167-179. doi: 10.2307/1510891

Ojemann, R. H. (1934). The reading ability of parents and factors associated with reading difficulty of parent education materials. *University of Iowa Studies: Child Welfare, 8*, 9-32.

Page, E. B. (1966, October). *Grading essays by computer*. Paper presented at the Progress Report Invitational Conference on Testing Problems, New York, NY.

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

Patty, W. W., & Painter, W. I. (1931). Technique for measuring the vocabulary burden of textbooks. *Journal of Educational Research, 24*, 127-134.

Poteet, J. A. (1979). Characteristics of written expression of learning disabled and non-learning disabled elementary school students. *Diagnostique, 4*(1), 60-74.

Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). *Stumping e-rater: Challenging the validity of automated essay scoring* (ETS Research Rep. No. RR-01-03; GRE NO; 98-08bP). Princeton, NJ: Educational Testing Service.

Quinlan, T. (2004). Speech recognition technology and students with writing difficulties: Improving fluency. *Journal of Educational Psychology, 96*(2), 337-346. doi: 10.1037/0022-0663.96.2.337

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct coverage of the e-rater scoring engine*. Princeton, NJ: Educational Testing Service.

Ransdell, S., & Levy, C. M. (1996). Working memory constraints on writing quality and fluency. In C. M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 93-105). Mahwah, NJ: Lawrence Erlbaum Associates.

Reppen, R. (2001). Register variation in student and adult speech and writing. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 187-199). Essex: Pearson ESL.

Salton, G., Yang, C., & Wong, A. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620. doi: 10.1145/361219.361220

Sheehan, K., Kostin, I., Futagi, Y., & Sabatini, J. (2007). *Reading level assessment for informational and literary texts*. Paper presented at the 29th Annual conference of the Cognitive Science Society, Nashville, TN.

Sheehan, K. M., Kostin, I., Deane, P., Hemat, R., Zuckerman, D., & Futagi, Y. (2006). *Inside sourcefinder: Predicting the acceptability status of candidate reading comprehension source documents* (ETS Research Rep. No. RR-06-24). Princeton, NJ: Education Testing Service.

Sheehan, K. M., Kostin, I., & Futagi, Y. (2007, October). *SourceFinder: A construct-driven approach for locating appropriately targeted teading comprehension source texts*. Paper presented at the SLaTE Workshop on Speech and Language Technology in Education ISCA Tutorial and Research Workshop, Farmington, PA

Spandel, V., & Stiggins, R. J. (1990). *Creating writers: Linking assessment and writing instruction* (2nd ed.). London: Longman.

Stewart, M. F., & Grobe, C. H. (1979). Syntactic maturity, mechanics of writing, and teachers' quality ratings. *Research in the Teaching of English, 13*, 207-215.

Strunk, W. (2000). *The elements of style*. Boston, MA: Allyn & Bacon.

Vogel, M., & Washburne, C. (1928). An objective method of determining grade placement of children's reading material. *Elementary School Journal, 28*(5), 373-381. doi: 10.1086/456072

Will, U., Nottbusch, G., & Weingarten, R. (2006). Linguistic units in word typing: Effects of word presentation modes and typing delay. *Written Language and Literacy, 9*(1), 153-176. doi: 10.1075/wll.9.1.10wil

Witte, S. P., Daly, J., & Cherry, R. (1986). Syntactic complexity and writing quality. In D. McQuade (Ed.), *The territory of language* (pp.150-164). Carbondale: Southern Illinois University Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, HI: University of Hawaii: Second Language Center.

## Notes

1 A T-unit is defined as an independent clause plus any associated dependent clauses. T-unit length is generally measured in words.

2 For instance, French (1961) identifies five such traits: ideas, form, flavor, mechanics, and working, which cover much the same territory as the 6-traits model.

3 In this context precision means the percent of responses labeled by the program that are labeled correctly. Natural language processing methodologies generally assess both precision and recall (the percentage of responses that should have been labeled that were in fact labeled correctly.) There is usually a tradeoff between precision and recall. In an AES context, it is more important to have high precision than high recall so as to avoid an excessive rate of false positives. ETS maintains a standard corpus of examples for each feature that it develops. Changes or modifications to the programs that identify features are tested against the original corpus and a second set of examples used for cross-validation.

4 Factor names given here are changed to be consistent with later analyses by Sheehan and her colleagues.

5 The last two factors in these analyses were excluded from later analyses as they revealed relatively little about grade level or genre characteristics of texts.

6 The subset covered grades 4, 6, 8, 10 and 12. Some grade 4 essays were eliminated from the original set due to difficulties with the distinction between persuasive and descriptive essay

prompts at that grade level, and corresponding essays at other grades were deleted to maintain the design in which each essay was administered at two adjacent grade levels with topics balanced across students and classes.

7 Each "Subset" represents a group of essays in which students encountered the conditions in the same order. For example, in one order, a student might be asked to compose: (i) a descriptive essay for an at-grade level prompt, (ii) a descriptive essay for a below-grade level prompt, (iii) a persuasive essay for an at-grade level prompt; and (iv) a persuasive essay for a below-grade level prompt.

8 Since only a subset of essays were scored by human raters in the Attali-Powers study, correlations for the latter are based on a smaller sample.

9 Human scores were available only for a subset of responses in the Attali-Powers data. They were collected as a cross-check on the Attali-Powers developmental scale and should be interpreted with some caution. We use them here primarily to explore the general direction of correlations and to suggest hypotheses for future research, as discussed in the body of the test.

10 However, in our data, there is actually a small increase in the proportion of grammatical errors in $10^{th}$ and $12^{th}$ grades. Analysis suggests that the errors in question tend to appear in relatively complex sentences, which are more frequent in the $10^{th}$ and $12^{th}$ grade essays.

11 While vocabulary features have the greatest weight, a number of syntactic features also play a role in these factors, reflecting the use of typically academic constructions (such as passives and use of relative clauses) on the one hand, or of typically oral patterns, on the other. See discussion below.

12 It should be noted that the fourth essay order involves a smaller set of essays, with more students missing, than the other three essay orders, and was administered at the end of the school year. However, the unrotated pattern matrices for each essay order assign strong, but opposite weights to orthographic accuracy on components 2 and 3. It is not unreasonable to consider that orthographic accuracy should reflect a lack of fluency at producing structured text, while correlating even more strongly, as the other three essay orders suggest, with a general low performance at following written conventions.

13 It was convenient to combine multiple analyses, since many of the features are highly correlated, and therefore competed with one another in a stepwise regression. The resulting equation represents a compromise between the predictions made by each analysis.

14 Criterion also has some capability to support peer review, though we have not as yet conducted any studies of how this capability is used in the classroom.