Research Report

Evaluating the Construct-Coverage of the e-rater® Scoring Engine

Thomas Quinlan

Derrick Higgins

Susanne Wolff

January 2009

ETS RR-09-01

Evaluating the Construct-Coverage of the e-rater® Scoring Engine

Thomas Quinlan, Derrick Higgins, and Susanne Wolff
ETS, Princeton, New Jersey

January 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This report evaluates the construct coverage of the e-rater® scoring engine. The matter of construct coverage depends on whether one defines *writing skill*, in terms of *process* or *product*. Originally, the e-rater engine consisted of a large set of components with a proven ability to predict human holistic scores. By organizing these capabilities into features, e-rater researchers organized the e-rater engine along the lines of trait scoring, which recognizes that essay quality has several dimensions. Some traits of essay quality cut across different methods for scoring essay quality, such as the rubrics employed by the GRE® and TOEFL® assessments, as well as the 6-trait scoring model. Factor analyses conducted by Attali and Powers (2008) suggest that e-rater features capture low-level aspects of essay quality, such as sentence complexity, vocabulary, and conventions. Future e-rater development should focus on (a) deepening and expanding coverage of the construct, such as by developing measures of essay content and organization, as well as on (b) addressing accuracy issues in existing features.

Key words: automated essay scoring, e-rater, writing assessment, writing skill

Acknowledgments

The final version of this report greatly benefited from the contributions of some key people. The authors would like thank Martin Chodorow and Jill Burstein for their very thorough comments on a previous version of this report, as well as Irene Kostin for her painstaking analysis of e-rater® performance statistics

Table of Contents

Introduction

In anticipation of adapting ETS's automated scoring capabilities to future, construct-driven writing assessments, it is imperative to gain a thorough understanding of the e-rater® scoring engine's construct coverage. Automated essay scoring (AES) affords the possibility of finer control in measuring the writing construct (Bennett, 2004). However, the realization of this possibility depends on reliable, construct-relevant measures of essay quality. Many of these automated measures may already be present in the e-rater engine. By evaluating the construct coverage of the e-rater system, we can identify e-rater measures with the greatest construct relevance. This report will: (a) provide a basic introduction to the e-rater scoring engine, (b) consider various definitions of writing competency, (c) evaluate the e-rater engine's construct coverage, and (d) assess the accuracy of the e-rater system's measurement capabilities. As such, it intends to address two questions regarding the research agenda for continuing capability development for the e-rater engine: (a) Which areas of the construct of writing should be prioritized for new feature development, and (b) which existing features should be prioritized as needing review or revision due to concerns over accuracy?

The e-rater Scoring Engine

ETS researchers developed the e-rater scoring engine to support the scoring of writing assessments with high-stakes outcomes (Attali & Burstein, 2006). In the year of its introduction, the e-rater engine assisted in scoring essays in the Graduate Management Admission Test (GMAT). In writing assessments that have high-stakes outcomes, essays are often scored by two trained human raters, with the final score being an average of the two scores. Double-scoring results in a more reliable final score (Breland, Camp, Jones, Morris, & Rock, 1987). In scoring the GMAT, the e-rater score took the place of the second human score. Originally, the e-rater scoring engine consisted of a large set of measures with a proven ability to predict human holistic scores. However, for the e-rater (v.2) engine, researchers aggregated this large set of measures into a smaller set of readily recognizable categories, called *features*. As the basic measures in the e-rater engine, the current set of eight features includes: grammar, usage, mechanics, style, organization, development, lexical complexity, and content (Attali & Burstein, 2006), though the specific features used in scoring depend on the model used (discussed in the following section). Some of these features (i.e., grammar, usage, mechanics, and style) are composed of multiple measures, so the term *microfeature* is used to distinguish any subfeature-

level measurement capability. For example, the subject–verb agreement microfeature contributes to the grammar feature. For a glossary of terms, see Appendix B.

E-rater scoring takes place in these major steps. First, its features and microfeatures measure many aspects of an essay, which are then aggregated into feature scores. Next, a weighted average of the feature scores is calculated to produce the final e-rater score. In calculating the final score, the weighting of feature scores is achieved by one of two methods, using either prompt-specific or generic models. Preliminary to actual scoring, ETS measurement specialists determine what method better suits the requirements of the assessment. With prompt-specific models, the weighting scheme is determined entirely empirically for individual prompts. ETS staff select a sample of essays that have been scored by two trained human raters (i.e., double-scored essays), process them through the e-rater scoring engine to obtain features scores for each essay, and then use regression analysis to determine the optimal weighting scheme that best predicts the average human score. With generic models, the weighting scheme is based on a pool of prompts and may be developed using empirical methods or expert opinion. In the latter case, writing experts decide on an a priori scheme, weighting each feature according to its construct relevance. The prompt-specific models often include the feature for measuring content (i.e., topic-specific vocabulary usage), whereas generic models never include this feature. There are also various hybrid methods of developing models that combine elements of generic and prompt-specific models. E-rater scores produced by both prompt-specific models and generic models have both shown to strongly predict holistic scores assigned by humans (Attali & Burstein, 2006).

Defining Writing Skill

Predicting human scores is one type of validity; construct validity is another. In modeling human holistic scores, predictors can have greater or lesser construct relevance. Consider, for example, essay length. Word count strongly predicts human holistic scores of impromptu essay examinations (Breland, Bonner, & Kubota, 1995). In spite of this relationship, essay length is a rather ambiguous indicator of essay quality. Certainly, a good essay requires a certain amount of development, and a longer document may tend to be qualitatively superior to a shorter document—other things being equal (i.e., ideas, organization, text structure, word choice, conventions, etc.). However, text length alone does not signify good-quality academic writing. Through a series of analyses, Chodorow and Burstein (2004) concluded that e-rater (v.1.0) scores

go beyond text length in accounting for the variance of human holistic scores. E-rater developers have aimed to develop scoring capabilities that model human scores while also having relevance to the writing construct.

The extent of the e-rater engine's construct relevance depends on how we define *writing skill*. Writing educators and researchers often talk about writing as either a *process* or a *product*. This controversy generally centers on how writing should be taught and assessed. In a product approach, writing skill is largely reflected in the final text produced, usually an essay. By contrast, the process approach recognizes that composing involves various types of problem solving (sometimes described at the level of *behaviors* or *cognitive processes*), operating recursively. In order to understand the e-rater scoring engine's coverage of the construct of writing competency, consideration of both product and process perspectives is useful.

A Process Approach to Writing Skill

Although now encompassing a variety of assumptions, the process approach was originally inspired by cognitive research. In their original model, Hayes and Flower (1980) described skilled writing in terms of three processes: planning (which generates and organizes content), translating (which converts ideas into words, transcribing words into text), and reviewing (which involves reading and editing). Their research falls into a long line of cognitive research aimed at distinguishing expert and novice performance in a domain. Hayes and Flower gathered data by observing skilled writers, who verbalized their thoughts (i.e., think-aloud protocols) during composing. By analyzing the protocols—as opposed to the texts—the authors could see that writers moved recursively through a range of problem solving. They concluded that this problem solving reflected three processes: planning, translating, and reviewing.

Bereiter and Scardamalia (1987) also analyzed think-aloud protocols to investigate how skilled writers differ from novices in terms of problem solving. They found that novice writers tended to take a simple, knowledge-telling approach to composing. In a knowledge-telling approach to writing, writers generate content through association, in which the topic, the discourse schema, and the developing text provide cues for generating content. The immature skills of novice writers restrict them to a knowledge-telling approach. In contrast, skilled writers may sometimes *problematize* a writing task, adopting an approach called *knowledge transforming*. More skillful writers often develop elaborate goals, particularly content and rhetorical goals, which require sophisticated problem solving. Skilled writers can move freely

between knowledge telling and knowledge transforming as needed, whereas the inefficient skills of novice writers may restrict them to knowledge telling.

Cognitive models of writing that define writing competency in terms of problem solving (e.g., Bereiter & Scardamalia, 1987; Hayes & Flower, 1980) tend to put the e-rater engine at a disadvantage. Writing researchers have long recognized that a written text very imperfectly reflects the problem solving of the writer who produced it. By reading an essay, one cannot determine whether the writer engaged in extensive drafting, evaluating, and revising (i.e., knowledge transforming) or simply dashed it off (i.e., knowledge telling). Since the e-rater engine was expressly designed to analyze text—specifically, a certain type of academic essay, written in response to an impromptu topic, under certain time restrictions—it cannot properly evaluate the writer's approach to problem solving.

Although the e-rater engine may not be able to distinguish between the problem solving of more- and less-skilled writers, it may measure aspects of basic writing skill. Although a written text does not provide a perfect picture of the writer's thinking, it does reveal something about the writer's ability to compose grammatical, well-punctuated sentences—the sine qua non of being a skillful writer. While Hayes and Flower (1980) conceived of skilled writing as a complex interleaving of processes, they also recognized that novice writing could involve simpler sequences. Accordingly, novice writing might consist mainly of translating, with little or no planning or reviewing. Skill in translating might be reasonably operationalized in terms of the speed and accuracy of composing basic sentences. Bereiter and Scardamalia's (1987) knowledge telling could also be conceptualized in terms of basic sentence composing. Assuming a knowledge-telling approach, there is reason to think that the e-rater engine can succeed in measuring basic writing skills (Attali & Powers, 2008).

A Product Approach to Writing Skill

In contrast to process approaches, educators and researchers have also taken a product approach to writing skill. Although educators, such as language arts teachers, now tend to emphasize the process of writing, writing assessments still typically focus on the final written product. The vast majority of secondary and post-secondary writing assessments involve composing an essay (a product). The validity of these essay examinations rest on an argument: An essay, written in response to an impromptu topic, constitutes a valid demonstration of writing

skill, and the quality of this essay provides a measure of writing skill. The quality of an essay is typically judged using either holistic or analytic trait methods.

In a product approach, writing skill equates to essay quality—with quality being very much in the eye of the beholder. A variety of aspects can affect the readability of an essay, and readers may notice one aspect more than another. For example, one reader may focus on the major points of an essay, and skim over minor inconsistencies, while another reader may be highly sensitive to mechanical errors. Hence, writing quality is inherently multidimensional (as well as inherently subjective). While both holistic and analytic trait methods recognize these various dimensions of writing quality, the two methods differ in how they operationalize the dimensions in scoring. In analytic trait scoring, quality dimensions are scored separately. This approach is useful when it is desirable to capture a student's strengths and weaknesses. Further, trait scores were designed to be instructive, with students using them for revising their essays (Spandel & Stiggins, 1990). By contrast, in holistic scoring, raters form an overall impression of an essay's quality, taking into account multiple dimensions at once, including clarity of ideas, text structure, and word choice. Consequently, the single holistic score tends to provide students with relatively less information about the relative strengths and weaknesses of their essay.

The development of the e-rater scoring has always focused on a written product, the essay. E-rater researchers developed automated capabilities (features and microfeatures) that provide good prediction of human judgments of essay quality. In order to improve face validity, they aggregated these capabilities into recognizable dimensions of essay quality (i.e., features), which resemble analytic traits. Then, as now, the Criterion^{SM} Online Writing Evaluation Service provides students with various categories of feedback on their essays. In these Criterion categories, e-rater researchers found a taxonomy for organizing the weighted feature model of the e-rater engine. Now, in scoring essays, the e-rater engine calculates scores for each feature, with the final score being a weighted combination of these feature scores. Thus, the e-rater engine already reflects a trait-scoring approach, whether superficially or functionally. The trait-scoring approach represents a sound product-oriented competency model that comports well with the e-rater scoring engine's measurement capabilities.

The 6-Trait Scoring Model

The Criterion service's categories of feedback mirror the analytical trait-scoring approach. Among approaches to analytical trait scoring, 6-trait scoring (Spandel & Stiggins,

1990) is perhaps the most well-known. Developed in collaboration with a group of classroom teachers, 6-trait scoring has garnered a following in the educational community. For example, the National Writing Project (NPW) recently used 6-trait scoring "to study the effectiveness of the writing project model and its impact on students in a range of contexts" (NWP, 2006, p. 1).

As originally developed, 6-trait scoring focuses on dimensions of students' written texts:

1. Ideas and content. The paper is clear, focused, and interesting. It holds the reader's attention. Relevant anecdotes and details enrich the central theme or story line.
2. Organization. The organization enhances and showcases the central idea or theme. The order, structure, or presentation is compelling and moves the reader through the text.
3. Voice. The writer speaks directly to the reader in a way that is individualistic, expressive, and engaging. Clearly, the writer is involved in the text and is writing to be read.
4. Word choice. Words convey the intended message in an interesting, precise, and natural way. The writing is full and rich, yet concise.
5. Sentence fluency. The writing has an easy flow and rhythm when read aloud. Sentences are well built, with consistently strong and varied structure that makes expressive oral reading easy and enjoyable.
6. Conventions. The writer demonstrates a good grasp of standard writing conventions (e.g., grammar, capitalization, punctuation, usage, spelling, paragraphing) and uses them effectively to enhance readability (Spandel & Stiggins, 1990).

Educators using the 6-trait scoring model now recognize a seventh trait for capturing the quality of the final presentation of student writing. Thus, the method is now commonly referred to as *6+1 trait scoring*.

The 6-trait scoring method was developed by teachers, for teachers. Teachers who use 6-trait scoring gather a wide range of information about their students' writing skills. Importantly, 6-trait scoring also provides students with information about the relative strengths and weaknesses of their written drafts, which the students can use for revising and editing. Thus, 6-trait scoring is primarily formative, providing students with potentially instructive feedback.

However, relative to holistic scoring, 6-trait scoring requires teachers to evaluate student writing more extensively, which may serve to dampen teachers' enthusiasm for it.

ETS was instrumental in developing methodologies for holistic scoring (Diederich, 1974), which was largely motivated by the need for improving the reliability of human scoring. Although much is known about the interrater reliability of holistic scoring, analytic trait-scoring approaches (e.g., 6-trait scoring) have not received the same research scrutiny. Thus, we know little about the relative stability of individual trait scores. The NWP used a modified version of 6+1 trait scoring (Bellamy, 2005) to rate student writing samples, as part of a large-scale evaluation of the impact of professional development at four National Writing Project (NWP) sites. After training the raters, which included calibrating to a Criterion level of performance, NWP researchers reported a rate of interrater agreement (either exact or adjacent) of between 90% and 95% (National Writing Project [NWP], 2006).

Analytic trait-scoring approaches (e.g., 6-trait scoring) highlight an important fact. In judging holistic quality, human readers are asked to consider a relatively few linguistic dimensions of an essay. If one surveys the holistic scoring rubrics of various writing assessments that employ the classic, persuasive essay task (see Appendix A), one notices the same few traits of essay quality appearing repeatedly. Typically, one or two traits specify high-level concerns, such as the quality and organization of ideas, while two or three other traits specify low-level issues, such as sentence fluency, word choice, and conventions. The appearance of these general traits, time and again, suggest that there is much consensus about the definition of essay quality. The definitions may differ slightly across contexts, but the underlying traits appear relatively stable.

The e-rater Scoring Engine's Coverage of Quality Traits

Given that essay quality has various dimensions, how well does the e-rater engine capture these dimensions? While each e-rater feature bears the name of a trait of essay quality, what can we infer about the breadth and depth of the coverage? The e-rater scoring engine has a somewhat hierarchical organization, with one or more measures (i.e., microfeatures) contributing to each feature score. Figure 1 illustrates how e-rater scoring is organized, with the measures (i.e., microfeatures) that underlie each feature. At first glance, one notes that the grammar, usage, mechanics, and style (GUMS) features have many more underlying microfeatures than the organization, development, lexical complexity, and topic-specific vocabulary usage features have. However, this does not necessitate a greater depth and breadth of coverage, since each

feature measures different things in different ways. For example, the grammar feature tallies certain types of errors, while lexical complexity measures average word length and word frequency. Thus, the number of underlying microfeatures does not necessarily indicate greater breadth or depth of construct coverage.

Another way to evaluate the construct coverage of the e-rater engine is to analyze the internal structure of the different features. Attali (2007) collected essay responses to the writing prompt of the *Test of English as a Foreign Language*™ computer-based test (TOEFL® CBT) impromptu argumentative essays. The author collected a broad sample, consisting of two essays from 5,006 examinees, from 31 different countries. Essays were processed using the e-rater scoring engine, with the extraction of individual feature scores, including development, organization, vocabulary, lexical complexity, style, mechanics, grammar, and usage. Confirmatory factor analysis yielded both two- and three-factor models, with the latter proving a better fit of the data. The three observed factors were relatively orthogonal, which the author interpreted as a discourse factor, a grammar factor, and a word usage factor.

Conducting a similar study, Attali and Powers (2008) investigated the factor structure underlying e-rater scores of essays written by primary and secondary students. The authors used the e-rater engine to analyze 30,600 essays, from 11,955 students of 261 schools in 527 classes, across grades 4, 6, 8, 10, and 12. Students wrote four essays, two expository and two persuasive. The e-rater engine measured the essays on seven features, including grammar, usage, mechanics, style, vocabulary, word length, and essay length. (Note: In their analyses, the authors replaced the e-rater engine's organization and development features with text length, after discovering that these two discourse features correlated strongly with the number of words in an essay.) A series of confirmatory factor analyses revealed that a 3-factor model provided the best fit for essays composed by students in the upper grades (grades 8, 10, and 12), and the loadings changed from one grade to the next. The authors interpreted these factors as relating to (a) fluency, (b) sentence-level conventions, and (c) word choice. Meanwhile, a 2-factor model provided a better fit for essays written by students in the lower grades (grades 4 and 6). The authors interpreted one factor as word choice, with the other representing a merging of fluency and sentence-level conventions.

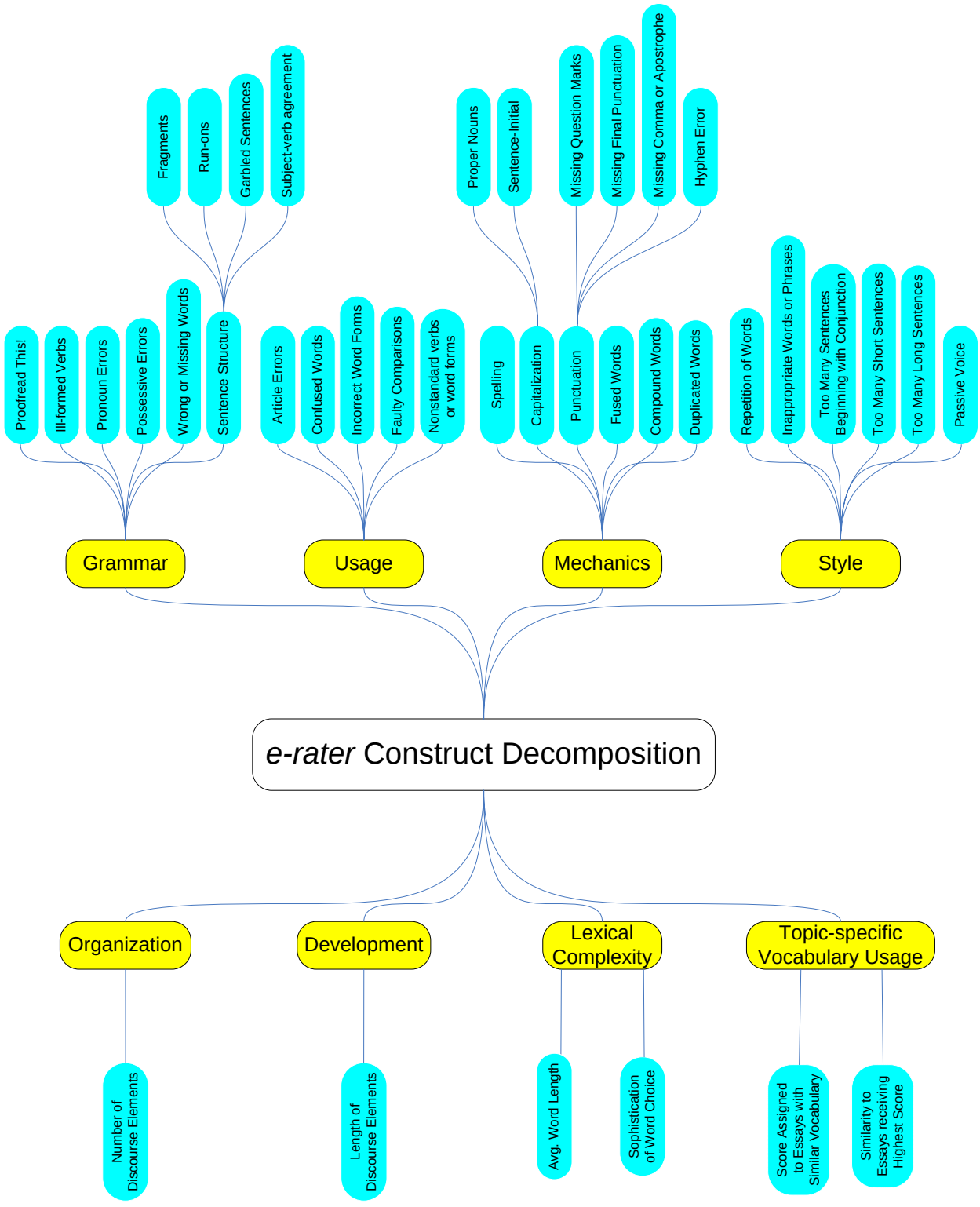

Figure 1. **The organization of e-rater scoring.**

The results of these two studies (Attali, 2007; Attali & Powers, 2008) converge in suggesting that the e-rater engine succeeds in measuring three aspects of writing skill, which can be loosely described as control of (a) grammar, conventions, and mechanics, (b) word usage, and (c) discourse structure and/or fluency. The relatively minor differences in results observed between the two studies may be explained by the developmental differences we might expect in these two divergent populations, L2 adults versus L1 students.

However, these studies also leave open a major question. To what extent might the observed factor structure be an artifact of the way the e-rater features and microfeatures are organized? When the e-rater scoring engine's measurement capabilities were organized into features, microfeatures were assigned to features on the basis of apparent similarity. Thus, it is not clear the extent to which the microfeatures within a feature interrelate statistically. Further, the results of these two studies suggest that that the underlying structure of e-rater measurement may diverge somewhat from the current feature scheme. This question is currently under investigation.

Improving Construct Relevance

Heretofore, e-rater development has aimed at optimizing the prediction of holistic scores assigned by human raters, especially for use in supporting the human scoring of high-stakes writing assessments. For ongoing development, improving construct relevance should be an overarching consideration. Toward this end, analytic trait scoring provides an apt theoretical framework for guiding future work. Assuming that all traits make some contribution to the overall quality of an essay, we would want to provide good coverage of each. In the development of e-rater features, the driving consideration has been increased prediction of human holistic scores. Yet, from a validity standpoint, it is not clear what this prediction means, since a feature can be either more or less construct-relevant. To address this problem, construct relevance should become an integral consideration in future e-rater development.

How might construct relevance become integral to the development of new e-rater features? Human scoring provides a model. Without sufficient training, humans will often assign different scores to particular essay. In training, a group of raters negotiate an interpretation of the rubric, which will often involve compromises between trait definition and implementation. In so doing, they operationalize the trait-scoring definitions in a construct-relevant way. In fact, the 6-trait scoring approach emerged from a very similar process, specifically, a group of teachers

discussing the important dimensions of student writing—a social process involving interpretation and negotiation (Spandel & Stiggins, 1990). E-rater development would benefit from a similar process of negotiation, to ensure that construct relevance remains central to development.

What would it mean to have construct relevance drive the development of new e-rater features? ETS researchers working on other applications of automated scoring have grappled with this question (Bennett, 2004; Bennett & Bejar, 1998). In order to realize the promise of AES to deliver finer control of the construct (Bennett, 2004), relevance to the construct would seemingly need to enter into the development process at multiple junctures. All types of assessment involve compromises. The development of new e-rater capabilities should involve a group consisting of natural language processing (NLP) researchers, writing educators, and writing assessment specialists. They should deliberate extensively to arrive at an optimal definition of the trait, that is both (a) amenable to NLP techniques and (b) maximally consistent with human interpretations. Spot-checks throughout development could help ensure that the implementation remains consistent with the original definition.

This development group should also establish the e-rater engine's research priorities. First, this group should identify the dimensions of essay quality most tractable for current NLP technologies. These dimensions may or may not conform exactly to the 6-trait scoring model but should represent a defensible interpretation of generally accepted dimensions of essay quality (e.g., as specified in scoring rubrics). Specific recommendations will require balancing the relevant issues, such as (a) the state of the art in NLP technology, (b) construct relevance, and (c) usefulness for different applications (i.e., essay scoring versus student feedback). Consequently, specific recommendations for extending construct coverage call for input from multiple perspectives. Even if one person could bring multiple perspectives to the process of developing new NLP capabilities, group deliberation (as described above) would seemingly provide checks and balances on development to help insure that new e-rater features remain relevant to the construct. As a starting point, they should consider the following issues:

- Voice. Measuring voice would require reliably identifying attempts to engage the reader—as well as evaluating the relative success of each attempt. Humans have difficulty assessing this trait. As a result, voice may be a relatively intractable construct, making it an unsuitable target for immediate development and so not an immediate research priority.

- Ideas and content. Although it may not be possible any time soon to automatically measure the quality of ideas in an essay, automatically assessing content is possible. The e-rater engine currently has two features for measuring content, which use content vector analysis (CVA) to measure the similarity of vocabulary between words in an essay and other texts (e.g., a prompt, a group of essays). However, CVA represents a rather basic approach to content evaluation, and has notable weaknesses (e.g., it fails to account for syntax and synonomy). To the extent that writing assessments often also evaluate content knowledge, we might want more-sophisticated measures of content.
- Organization. Strictly speaking, this trait evaluates the flow of ideas in an essay—something current AES capabilities have difficulty capturing. However, it should be possible to evaluate whether an essay is generally well structured. Producing well-structured text is an important aspect of writing skill, as the writer must learn to build up words into sentences, sentences into paragraphs, and paragraphs into a coherent document. The e-rater scoring engine currently has two features (Organization and Development) originally designed to identify the parts of an essay, as distinct units of discourse (i.e., introduction, thesis statement, main idea, supporting idea, and conclusion). These features were revised, so that they currently measure the number and average length of discourse units, which collectively correlate strongly with text length. New e-rater features might evaluate organization in terms of aptness of essay paragraph breaks and overall essay coherence.
- Word choice. AES capabilities may have difficulty assessing the extent to which the words in an essay convey meaning to a reader. The e-rater engine currently has two features that measure the relative sophistication of word choice. One calculates the frequency of words used in an essay, relative to their occurrence in a very large corpus of popular publications. The other calculates the median number of charters per word in an essay. To the extent that word choice relates closely to content, current research into automatically scoring the content of constructed responses may eventually contribute to measuring word choice.

- Sentence fluency. Again, AES may have difficulty capturing the strict definition of this trait (i.e., the readable flow of sentences in an essay). However, the e-rater engine currently provides some sentence-level measurement of an essay, corresponding to the writer's ability to construct grammatical, properly punctuated sentences. The ability to properly punctuate sentences, so as to avoid making fragment and run-on sentences, represents an important aspect of basic writing skill. Robust measures of sentence complexity could capture this trait.
- Conventions. Most of the e-rater engine's current microfeatures measure aspects of conventions. Therefore, we might conclude that the e-rater engine has very good coverage of this trait. Development efforts should focus on microfeatures with possible weak accuracy. See the discussion that follows.

Whether taking a process (cognitive) or product (trait-scoring) approach, the e-rater scoring engine appears to capture aspects of basic text production (i.e., the skills responsible for arranging words into grammatically correct sentences). A trait-scoring approach (e.g., 6-trait scoring) offers the greatest opportunity for fine-tuned control of the construct, since it defines quality in terms of measurable aspects of an essay, unlike cognitive approaches that define *competency* in terms of processes. Further, the trait-scoring approach has an additional advantage: If e-rater features provide sound measures of traits of essay quality, then those trait measures can be easily repurposed for other assessment purposes, besides holistic scoring of high-stakes writing assessments. Therefore, trait scoring provides a useful theoretical and practical framework to guide the development of new scoring features. Embracing a trait-scoring approach will seemingly necessitate investigations to better understand the performance of each e-rater feature, vis-à-vis its ostensible trait.

Accuracy of e-rater Microfeatures

The construct coverage of the e-rater scoring engine ultimately depends on the measurement accuracy of each microfeature. Although the importance of accuracy is obvious, its relative impact varies according to application. In essay scoring, accuracy issues may be less critical, since we might expect the aggregation of microfeatures into feature scores and the statistical modeling of human scores would serve to make e-rater scoring relatively robust. Poor accuracy in one or two microfeatures may have little impact on the overall e-rater score and its

correlation to human scores. In contrast, in the Criterion writing environment, accuracy problems (i.e., false-positive errors) mean that students get erroneous feedback on their essays. We currently have few insights into how accuracy problems may or may not disrupt students in the act of writing. We can safely assume that the importance of accuracy varies, depending on the specific application of the e-rater scoring engine, whether for providing student feedback or predicting human holistic scores.

Accuracy is an essential part of the evidentiary argument supporting e-rater scoring. Accuracy has two dimensions, precision and recall. Precision refers to the rate of false-positives, in which human–system agreement is based on system decisions; this measure answers the question of how often the human and system agree on system assignments. *Recall* refers to the rate of false–negatives, in which human–system agreement is based on human assignment; this measure answers the question of how often the human and system agree on human assignment. The development of most types of automated scoring requires a compromise between precision and recall. E-rater developers aim for high precision (i.e., 80%), since false–positives are much more disruptive than false–negatives in both essay scoring and Criterion scoring. However, once installed in the e-rater engine and deployed in the real world of scoring actual essays, this relatively high accuracy may or may not hold, since the performance of one microfeature may influence the performance of another. Thus, e-rater performance should be considered comprehensively, at three levels: microfeature, feature, and system.

Although the importance of accuracy is self-evident, evaluating it poses some surprising methodological challenges. New NLP capabilities are typically evaluated against some human annotated corpora. Researchers speak of developing a gold standard corpus that is (a) highly representative of the target domain (e.g., student essays) and (b) very accurately annotated. To ask a seemingly obvious question: Why human annotation? To the extent that AES systems (like the e-rater scoring engine) aim to model human scoring of essays, human performance would seem to be a sound basis of comparison. However, this is questionable, since even trained human raters often disagree in the scores they assign to a particular essay (Breland et al., 1987). The history of writing assessment can be viewed as an ongoing struggle to attain acceptable levels of interrater agreement (Eliott, 2005). In fact, ETS researchers (i.e., Diederich, 1974) developed holistic scoring methodologies for scoring student essays largely to deliver acceptable levels of interrater reliability. Human raters make an overall judgment of essay quality, and training

involves meeting strict levels of agreement among raters. Reliability is further improved by double-scoring, with the scores of two raters combined to yield a more-stable, final essay score. These issues equally apply to annotation, which resembles scoring in many ways. Consequently, the same rigorous training may be necessary for annotators to establish acceptable levels reliability.

However, the reliability of annotation is further undermined by the ambiguity of certain linguistic phenomena in an essay. Whereas a spelling error may be identified with a high degree of certainty, it may not be possible to definitively identify the precise location and nature of other types of errors. To illustrate, consider the following sentence from an essay written for a high-stakes writing assessment: "In consion [sic], for some reasons, museum, particularly known travel place, get on many people." (example from Tetreault & Chodorow, 2008). Most native readers of English might agree that this sentence violates patterns of written English. However, without knowing the writer's intended ideas, the reader cannot identify the precise location and nature of the violation. Given this relative indeterminacy, human readers may often disagree because certain errors may license multiple error categories.

In theory, a gold standard corpus would provide a sound basis of comparison for evaluating an NLP capability. In actuality, a gold (or even silver) standard corpus may be difficult to attain in the field of AES. The weaknesses of human annotation are not well understood. As a basis for evaluating any AES system, they are a cause of concern.

Evaluating Accuracy via Interrater Agreement

The original purpose of the following investigation was to evaluate the performance of the e-rater engine, by comparing the Criterion writing-evaluation's grammar-checking capabilities to the capabilities of a popular word-processing program. This investigation illustrates some of the challenges of evaluating an AES system using a human-annotated corpus.

Method

Burstein, Chodorow, and Higgins (2007) conducted a pilot investigation to evaluate the accuracy of Criterion feedback, which is produced by the e-rater engine. In so doing, the authors evaluated the accuracy of 28 e-rater grammar, usage, and mechanics microfeatures. The authors randomly selected 2,400 sentence strings from the Criterion database of student essay submissions. Sentences were extracted automatically, using a sentence tokenization program that

finds sentences in a text. In this extracted corpus, some 62 non-sentences were incorrectly identified as sentences and were removed to the final corpus. The e-rater scoring engine was run on the sentence corpus and the flagged errors from the set of 28 error types. The authors proceeded with two different analyses. In the error-verification task, a human rater was given a subset of the corpus consisting of sentences with e-rater-identified errors. The rater read each sentence in the corpus, indicating whether or not she or he agreed with the error label assigned by the e-rater scoring engine. In the other analysis (the comprehensive error annotation task), four human annotators were trained to identify the 28 error types examined in the study. During training, the annotators were given opportunities to practice and build consistency between raters. Researchers gave a subset of 600 unlabeled sentences to each annotator, asking him or her to identify errors and label them according to error type.

Results

Problems with the annotation methodology became apparent during analyses, which tended to undermine the interpretability of the results. Consequently, we will only summarize those results, rather than report them in detail. The two analyses yielded somewhat divergent results. The error-verification task showed fairly strong scores for precision, with only one microfeature showing somewhat low precision (i.e., [202] Missing/extra article). However, it should be noted that this error type requires context, and so accuracy is lost when the sentence is evaluated outside the context of the full essay. In contrast, the comprehensive error-annotation task revealed very low precision on several microfeatures. A comparison of the two analyses suggested one likely explanation for the divergence. On the one hand, error verification may result in inflated scores. When presented with a supposed error, a human may tend to agree with an already assigned category, which will tend to bias the error-verification results in favor of agreement. On the other hand, the comprehensive error-annotation task may tend to yield depressed agreement scores. Since certain errors may legitimately fit more than one category, they may be miscategorized unless human annotators are given highly explicit decision rules. Even then, ambiguity of error type will tend to make exact human/system agreement less likely. These considerations suggest how the evaluation of microfeature accuracy may be influenced by characteristics of the human annotation.

Burstein et al.'s (2007) investigation highlights the difficulties involved in developing a gold standard corpus, as a basis for evaluating e-rater accuracy. Although some microfeatures

show low accuracy scores, these results may reflect either a deficiency in the microfeature, the nature of the corpus, or problems with human annotation. In particular, we have evidence that erroneous annotations by humans may have been prevalent for some of the microfeatures, which would have tended to depress accuracy scores in a cascading fashion—in some cases, dramatically. An ETS NLP research consultant who has developed many of the grammatical error microfeatures conducted an informal investigation of the annotated corpus, inspecting the errors identified by the human annotators. She found many instances of errors that were either (a) falsely identified or (b) miscategorized. Her investigation for the wrong word form microfeature, in the following discussion, is illustrative.

In the original analysis (Burstein et al., 2007), the human annotator flagged 233 wrong word form errors, while the e-rater system flagged none. If we broaden the agreement criteria to count instances where the e-rater system identified an error but classified it differently, we observed 40 instances of agreement. In other words, out of the 233 wrong word form errors identified by the human annotator, the e-rater engine identified 40 as some category of error. These results suggest that this microfeature performs rather poorly, by detecting relatively few instances of an apparently common error. Upon further examination, it turned out that, of the 231 instances incorrectly identified by the human annotator, 51 were falsely identified and 180 were miscategorized. In fact, wrong word form errors were actually quite infrequent, with only 2 out of the 233 errors labeled correctly.

This example highlights the potential problems of judging the accuracy of e-rater microfeatures against the standard of human annotation. In order for agreement between a system and humans to be a meaningful metric, (a) the labeling protocol must be well matched with the system capability, and (b) human annotation must be highly reliable. In this investigation, employing a second annotator would have provided an indication of the relative reliability of annotation, by which reliability problems might have been more readily apparent and perhaps addressed.

Still, the weaknesses of human evaluation potentially undermine the evaluation of accuracy in at least two ways. First, the relative frequency of a certain type of error has a major bearing on accuracy. When evaluating the accuracy of a microfeature, such as wrong word form, missing two errors is much less of a concern than missing 233. However, because of suspected problems in the human annotation, the relative frequency of certain error types cannot

confidently gauged. Second, given the nature of reading comprehension, it may not be possible to identify the precise location of an error in a sentence. Sometimes humans and the e-rater system differed on the precise location of an error, which accounts for some of this disagreement. That leaves us to suspect that some microfeatures are underperforming, without knowing for certain. Thus, when a microfeature fails to detect anything, we are left to speculate about the reasons, whether it may be due to the nature of the corpus, problems with human annotation, or actual insensitivity of the microfeature.

Evaluating Accuracy via Performance Statistics

While agreement with human annotation provides a convenient, face-valid metric for evaluating accuracy, in certain cases, the weak reliability of human annotation undermines the usefulness of this method. At minimum, creating a gold standard corpus would be extremely difficult. Consequently, alternative methods should be developed to provide additional sources of information about e-rater accuracy. One alternative approach might be to analyze how microfeatures perform, to assess whether performance conforms or deviates from expected patterns. Given a large, representative corpus of student essays, simple performance statistics should allow us to test some modest hypotheses about how microfeatures perform, in terms of frequency and distribution. For example, since microfeatures are intended to measure writing skill, most should be negatively related to student grade. Further, we should observe weak or no correlations between many other microfeatures. Since new e-rater microfeatures must demonstrate an 80% level of precision (i.e., less than 20% rate of false-positives) before they are approved for integration into the e-rater scoring engine, we might assume that they are performing well—unless we have evidence to the contrary.

Method

The following analysis presents an alternative approach to evaluating the accuracy. Instead of evaluating the performance of e-rater microfeatures against human annotation, this approach searches for statistical anomalies in how microfeatures perform on a large corpus of student essays. In the development of the e-rater scoring engine, individual microfeatures were designed to (a) measure construct-relevant aspects of essay quality and (b) contribute to the prediction of holistic scores assigned by human raters. By using simple descriptive statistics, we sought to identify e-rater microfeatures that were performing in unexpected ways.

For this analysis, we used a corpus of student essays developed by Attali and Powers (2008). The authors collected essays from 11,955 students, of 261 schools, in 527 classes, across grades 4, 6, 8, 10, and 12. Students wrote four essays, two expository and two persuasive. The Attali–Powers corpus was divided into four subsets, based on the order in which participants encountered the four essay tasks. 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 was processed using e-rater to produce a dataset of frequency information for individual microfeatures.

A series of statistical analyses were conducted to calculate frequency of individual microfeatures and correlations among them.

Results

Frequency analyses. In calculating frequency, three e-rater microfeatures were found to have zero or very few cases in the Attali/Powers dataset (N = 30,599). Two microfeatures had no cases: Run-on sentence error and Preposition error. One microfeature had only 10 cases: Wrong Word Form, which identifies a verb used in place of a noun.

The frequency of two other microfeatures appears to be errors of gross overestimation: Missing Initial Capitalization and Missing Final Punctuation microfeatures. These microfeatures correlated strongly with the number of line breaks. Informal examination revealed that the insertion of manual line breaks could trigger these microfeatures.

When a student pastes in an essay from Notepad (or another word processor), extra line breaks may be inadvertently inserted.

The spelling microfeature apparently overestimates spelling errors by counting proper nouns of various sorts. For example, in two essays, correctly spelled words having to do with "SpongeBob SquarePants" were coded as spelling errors. (Apparently, Microsoft Office Word also does not have many common proper nouns in its spellchecker.)

Correlational analyses. Although the sample as a whole has the expected correlation between grade level and sentence length (i.e., +.429), with longer sentences being associated with higher grade levels, there are several fourth-grade essays whose average sentence length is more than 2.5 standard deviations above the mean average sentence length for the entire sample. After examining three fourth-grade essays with average sentence lengths of more than 4.75 standard deviations above the mean for the entire sample, we observed a frequent absence of

terminal punctuation, resulting in run-on sentences. In effect, their essays were one long sentence. As discussed previously, the feature that is supposed to detect run-on sentences is, in fact, detecting no cases of run-on sentences.

Correlational analysis revealed that several e-rater features were statistically related, when we would have expected them to be relatively independent. First, fused word and spelling were strongly correlated. Informal examination of the data suggests that these two microfeatures are correlated because fused errors are being counted twice, once by each microfeature. Second, fragment and too -many-short-sentences appeared strongly correlated. Examination of the data suggests that fragments are being counted twice.

The passive voice feature correlated positively with the overall style rating, when the direction of the correlation should be negative. Writing handbooks often advise writers to use the active voice (Strunk, 2000), and the passive voice microfeature was developed to measure this supposed violation of style. However, the passive voice is often warranted in academic writing and so cannot be rightly considered a violation of style. Notably, the passive voice only detects a certain type of construction, by-passives.

The correlational analysis also revealed a surprising relationship between human holistic scores and feature scores for grammar, usage, and mechanics. When the e-rater engine identifies no cases of errors within a feature category, the feature score correlates strongly with the human score. For essays with one or more GUMS101-109 errors, the correlation is $r = .49$. In contrast, for essays with no GUMS101-109 errors, the correlation is $r = .86$. Essay length appears to be the mediating variable. The four GUMS features are calculated in a similar manner: by summing the total errors within each feature category, dividing by the number of words in the essay, and then taking the negative square root. When no cases are detected for a particular feature, then the e-rater system uses the formula -sqrt(1/words). Thus, when no errors are identified within an error category, the feature score becomes a transformed measure of text length—always a strong predictor of human quality scores. For essays in the Attali and Powers (2008) corpus, this (no errors identified) occurred relatively frequently (e.g., 43% for grammar and 27% for usage).

Conclusions About Accuracy

Currently, there is no gold standard method for evaluating the accuracy of e-rater features and microfeatures. Putting aside the question about whether developing a gold standard corpus is possible, we are clearly years away from having one. Consequently, in the second evaluation, we

explored an alternative method of evaluating e-rater accuracy. The e-rater scoring engine was used to score a large corpus of student essays at the microfeature level, with descriptive statistics used to identify unexpected frequencies and correlations. The results of the latter investigation provide readily interpretable evidence about the performance of e-rater microfeatures, by which we can identify those of greater and lesser concern. These judgments are summarized in Figure 2, in which *high concern* and *moderate concern* are indicated, respectively, by red and yellow shading.

High Concern

The ability to properly structure sentences is an important, foundational writing skill. In the primary grades, children learn to structure simple sentences by dividing sentences with proper punctuation. The correlational analysis of sentence length suggests that younger students may sometimes fail to punctuate sentences properly, suggesting that they may not have mastered an ability to construct simple sentences. Two microfeatures ostensibly measure this ability—run-on sentence and fragment, both of which have apparent problems. First, the run-on sentence microfeature does not appear to be functioning. An investigation should be conducted to determine whether this is an isolated or a general problem. Second, it appears that the fragment and the too many-short-sentences microfeatures are double-counting errors due to sentence fragments. From a construct perspective, reliably detecting sentence fragments is very important. In contrast, having too many short sentences does not necessarily, in itself, detract from the perceived quality of an essay. A defensible remedy to this apparent problem may be to simply remove the too-many-short-sentences microfeature from the e-rater engine.

When students use the Criterion service, they often cut and paste documents from their word processors. With this practice, the Criterion service appears to interpret line breaks as missing initial capitalization and final punctuation errors. This problem would tend to inflate the mechanics feature score.

Moderate Concern

The spelling microfeature is implicated in two apparent problems. First, the spelling microfeature tends to inaccurately identify many popular proper nouns as errors. Whenever a student writes about a cartoon character or a favorite sports player, it may trigger a spelling error. The spelling microfeature uses the Aspell open-source spellchecker (http://aspell.net/). It should be possible to create a list of proper nouns, as exceptions to spelling errors. Second, the fused

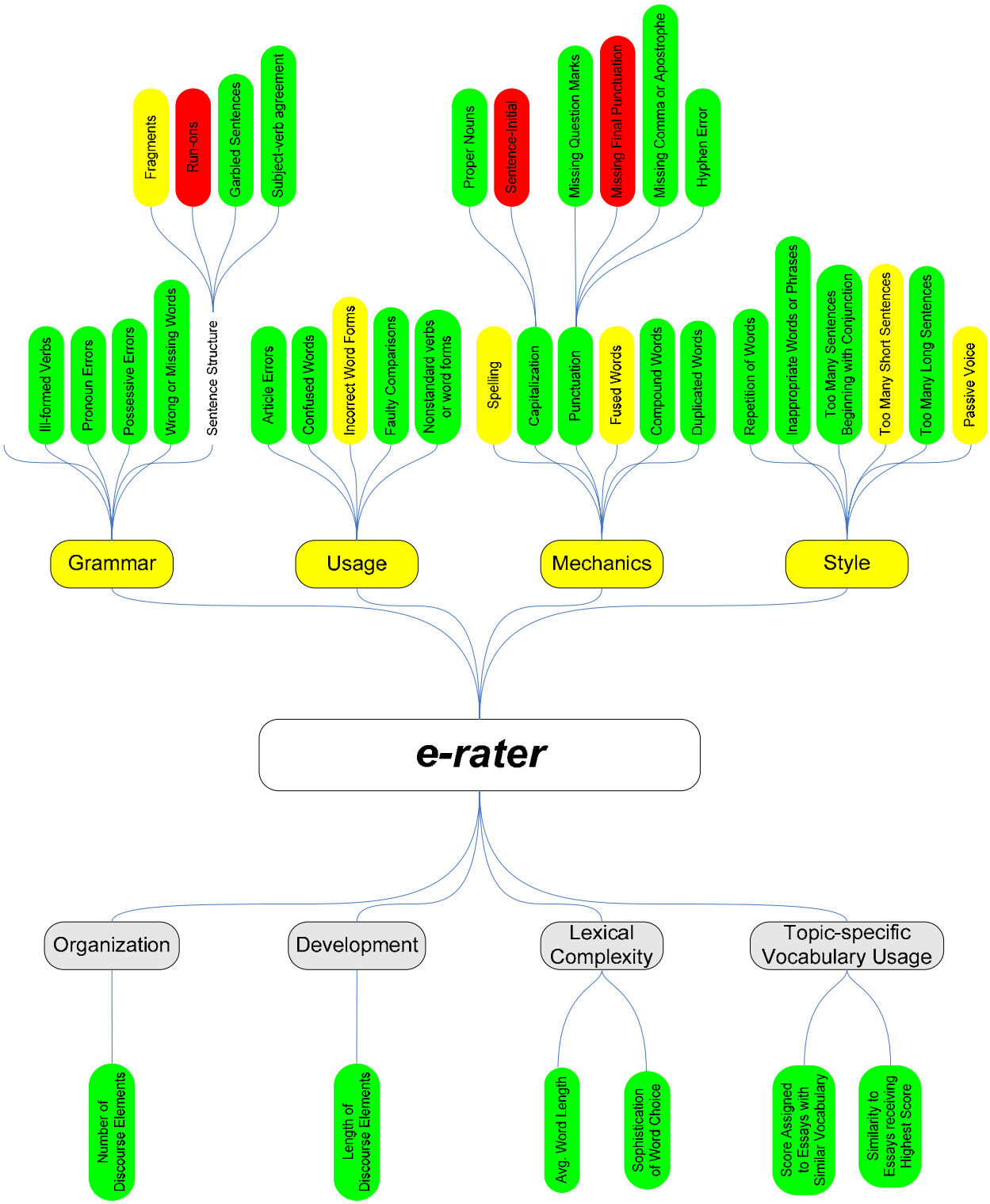

Figure 2. **Suspected accuracy issues in e-rater features and microfeatures.**

word microfeature correlated strongly with the spelling microfeature, suggesting that this type of error is being counted twice. Since the spelling microfeature apparently detects this type of error, this problem might be resolved by removing the fused word microfeature.

The passive voice microfeature apparently only identifies the *by-passives* and does not identify the agentless passives, which are the more common form. By-passives can be an appropriate verb construction in academic types of writing, which might help explain the observed positive correlation. Although this feature may contribute to prediction, it is not defensible from a construct perspective.

Text length appears to be frequently entering indirectly into e-rater models. When no errors are detected for a GUMS category, the feature score becomes a transformed measure of text length. Although the current algorithm provides a reasonable solution to the statistical problem of data sparsity—while also tending to boost the prediction of human holistic scores—it is not very defensible from a construct perspective. The formula used for aggregating feature scores should be changed so that feature scores do not become transformed measures of text length.

If one cannot rule out the influence of extraneous factors (e.g., characteristics of the corpus, appropriateness of the analysis, etc.), any evaluation remains problematic. As illustrated by the Burstein et al. (2007) study, the composition of the corpus, methods for human annotating, and type of analysis can each influence the validity of an evaluation. Developing a gold standard corpus would require addressing a host of questions—about the writer (grade level, language learner), the type of composition (narrative versus persuasive), the level of sampling (sentence versus continuous text), and the stage of writing process (single draft versus revised). Assuming a gold standard corpus were feasible, developing it would require considerable time and resources. This report demonstrates the potential value of alternative methods for evaluating e-rater performance.

Overall Conclusion

How well does the e-rater scoring engine measure writing skill? In the preceding sections, we have established that the e-rater system addresses some traits of essay quality, with some features/microfeatures performing better than others. Construct relevance and performance both have a major bearing on construct coverage, which gives rise to some related questions. How closely aligned are e-rater features to traits of essay quality? Do the features measure deeper or shallower linguistic aspects of the essay? How accurate are the measurements? Based

on the foregoing discussion, it is possible to make some preliminary judgments about these questions. Table 1 illustrates how e-rater features map to common traits of quality, as defined by scoring rubrics of the 6-trait model, *Graduate Record Examinations*® (GRE®), and TOEFL. It includes the first author's judgment about the e-rater engine's alignment to the construct, as well the relative depth of measurement. Lastly, Table 1 indicates whether or not we have evidence of accuracy issues for a particular e-rater feature. In sum, Table 1 indicates that the e-rater engine has some coverage of high-level aspects of essay quality, such as ideas and organization, and somewhat extensive coverage of low-level aspects (e.g., word choice, grammar, and conventions), with some accuracy issues in the latter.

Whether defined in terms of process or product, the e-rater scoring engine provides partial coverage of the construct, with the majority of measurement capturing the low-level aspects of essay quality that reflect basic writing skills. Future development should address suspected accuracy issues, then turn toward deepening and extending the coverage of traits of essay quality (e.g., 6-trait scoring). While feature development necessarily involves operationalizing trait constructs in a way amenable to NLP techniques, some interpretations may be more construct-relevant than others. To the extent that an essay reflects skills of the writer, new features should prove effective in identifying and measuring traits of essay quality by interpreting the construct in a manner consistent with human interpretations.

Putting the e-rater scoring engine on more solid trait-scoring footing will pose new challenges. Prediction of human scores has provided a clear target for e-rater development, but it also lacks transparency. Further, statistical relationships can be spurious or mediated by unforeseen variables. For example, the strong correlation between holistic quality and text length may be considered construct-irrelevant, apart from the mediation of variables such as organization, development, and sentence fluency. Orienting e-rater development toward trait scoring will seemingly require a new approach, with a systematic consideration of construct relevance throughout the development process.

Table 1

Overall Evaluation of Construct Coverage

e-rater feature	6-trait scoring model (Spandel & Stiggins, 1984)	GRE persuasive rubric (maximum score)	TOEFL independent rubric (maximum score)	Aligned to construct	Depth of measure	Accuracy issues?
n/a	Ideas and content	Presents an insightful position on the issue	Effectively addresses the topic and the task	n/a		
Development		Develops the position with compelling reasons and/or persuasive examples	Uses clearly appropriate explanations, exemplifications, and/or details	Sufficient	Minimal	No
n/a	Voice	n/a	n/a	n/a		
Organization	Organization	Sustains a well-focused, well-organized analysis, connecting ideas logically	Is well-organized and well-developed	Sufficient	Minimal	No
n/a			Displays unity, progression, and coherence			n/a
Word choice	Word choice	Expresses ideas fluently and precisely, using effective vocabulary and sentence variety	Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors.	Sufficient	Sufficient	No
n/a	Sentence fluency			n/a	n/a	n/a
Grammar usage mechanics style	Conventions	And demonstrates facility with the conventions (i.e., grammar, usage, and mechanics) of standard written English but may have minor errors.		Good	Good	Yes

References

Attali, Y. (2007). *Construct validity of* e-rater *in scoring TOEFL essays* (ETS Research Rep. No RR-07-21). Princeton, NJ: ETS.

Attali, Y., & Burstein, J. (2006). 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* (ETS Research Rep. No RR-08-19). Princeton, NJ: ETS.

Bellamy, P. C. (Ed.). (2005). *Seeing with new eyes*. Portland, OR: Northwest Regional Educational Laboratory.

Bennett, R. E. (2004). *Moving the field forward: Some thoughts on validity and automated scoring* (ETS Research Rep. No RM-04-01). Princeton, NJ: ETS.

Bennett, R. E., & Bejar, I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9–17.

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

Breland, H. M., Bonner, M. W., & Kubota, M. Y. (1995). *Factors in performance on brief, impromptu essay examinations* (College Board Rep. No. 95-04). New York: College Entrance Examination Board.

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

Burstein, J., Chodorow, M., & Higgins, D. (2007). *Evaluation of* Criterion *feedback codes for sentence checking in FMI's ProofWriter.* Unpublished manuscript.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating* e-rater's *performance on TOEFL essays* (ETS Research Rep. No. RR-04-04; TOEFL Report No. RR-73). Princeton, NJ: ETS.

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

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

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.

National Writing Project. (2006). *Local site research initiative report (Cohort II) 2004–2005*. Berkeley, CA: Author.

Spandel, V., & Stiggins, R. J. (1990). *Creating writers: Linking assessment and writing instruction*. New York: Longman.

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

Tetreault, J. R., & Chodorow, M. (2008). *The ups and downs of preposition error detection.* Paper presented at the annual meeting of COLING, Manchester, UK.

Appendix A

Scoring Guides for the Graduate Record Examinations® (GRE®), Test of English as a Foreign Language™ (TOEFL®), SAT®; and National Assessment of Educational Progress (NAEP)

Score	GRE scoring guide	TOEFL	SAT	NAEP
6	A 6 paper presents a cogent, well-articulated analysis of the complexities of the issue and conveys meaning skillfully. A typical paper in this category: • presents an insightful position on the issue • develops the position with compelling reasons and/or persuasive examples • sustains a well-focused, well-organized analysis, connecting ideas logically • expresses ideas fluently and precisely, using effective vocabulary and sentence variety • demonstrates facility with the conventions (i.e., grammar, usage, and mechanics) of standard written English but may have minor errors.	(Score range: 0-5) An essay at this level largely accomplishes all of the following: • effectively addresses the topic and task • is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details • displays unity, progression, and coherence • displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors	An essay in this category demonstrates clear and consistent mastery, although it may have a few minor errors. A typical essay • 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 • demonstrates meaningful variety in sentence structure • is free of most errors in grammar, usage, and mechanics	Excellent response • takes a clear position and supports it consistently with well-chosen reasons and/or examples; may use persuasive strategy to convey an argument. • is focused and well organized, with effective use of transitions. • consistently exhibits variety in sentence structure and precision in word choice. • errors in grammar, spelling, and punctuation are few and do not interfere with understanding.
5	A 5 paper presents a generally thoughtful, well-developed analysis of the complexities of the issue and conveys meaning clearly. A typical paper in this category • presents a well-considered position on the issue • develops the position with logically sound reasons and/or well-chosen examples • maintains focus and is generally well organized, connecting ideas appropriately • expresses ideas clearly and well, using appropriate vocabulary and sentence variety • demonstrates facility with the conventions of standard written English but may have minor errors.		An essay in this category 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 • demonstrates variety in sentence structure • is generally free of most errors in grammar, usage, and mechanics	Skillful response • takes a clear position and supports it with pertinent reasons and/or examples through much of the response. • is well organized, but may lack some transitions. • exhibits some variety in sentence structure and uses good word choice; occasionally, words may be used inaccurately. • errors in grammar, spelling, and punctuation do not interfere with understanding.

Score	GRE scoring guide	TOEFL	SAT	NAEP
4	A 4 paper presents a competent analysis of the issue and conveys meaning adequately. A typical paper in this category • presents a clear position on the issue • develops the position on the issue with relevant reasons and/or examples • is adequately focused and organized • expresses ideas with reasonable clarity • generally demonstrates control of the conventions of standard written English but may have some errors.	An essay at this level largely accomplishes all of the following: • addresses the topic and task well, though some points may not be fully elaborated • is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details • displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections • displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, • though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning	An essay in this category 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	Sufficient response • takes a clear position and supports it with some pertinent reasons and/or examples; there is some development. • is generally organized, but has few or no transitions among parts. • sentence structure may be simple and unvaried; word choice is mostly accurate. • errors in grammar, spelling, and punctuation do not interfere with understanding.

Score	GRE scoring guide	TOEFL	SAT	NAEP
3	A 3 paper demonstrates some competence in its analysis of the issue and in conveying meaning but is obviously flawed. A typical paper in this category exhibits one or more of the following characteristics: • is vague or limited in presenting or developing a position on the issue • is weak in the use of relevant reasons or examples • is poorly focused and/or poorly organized • presents problems in language and sentence structure that result in a lack of clarity • contains occasional major errors or frequent minor errors in grammar, usage or mechanics that can interfere with meaning.	An essay at this level is marked by one or more of the following: • addresses the topic and task using somewhat developed explanations, exemplifications, and/or details • displays unity, progression, and coherence, though connection of ideas may be occasionally obscured • may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of • clarity and occasionally obscure meaning • may display accurate, but limited range of syntactic structures and vocabulary	An essay in this category 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	Uneven response (may be characterized by one or more of the following) • takes a position and provides uneven support; may lack development in parts or be repetitive OR response is no more than a well-written beginning. • is organized in parts of the response; other parts are disjointed and/or lack transitions. • exhibits uneven control over sentence boundaries and sentence structure; may exhibit some inaccurate word choices. • errors in grammar, spelling, and punctuation sometimes interfere with understanding.
2	A 2 paper demonstrates serious weaknesses in analytical writing. A typical paper in this category exhibits one or more of the following characteristics: • is unclear or seriously limited in presenting or developing a position on the issue • provides few, if any, relevant reasons or examples • is unfocused and/or disorganized • presents serious problems in the use of language and sentence structure that frequently interfere with meaning • contains serious errors in grammar, usage, or mechanics that frequently obscure meaning.	An essay at this level may reveal one or more of the following weaknesses: • limited development in response to the topic and task • inadequate organization or connection of ideas • inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task • a noticeably inappropriate choice of words or word forms	An essay in this category 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 demonstrates 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	Insufficient response (may be characterized by one or more of the following) • takes a position but response is very undeveloped. • is disorganized or unfocused in much of the response OR clear but very brief. • minimal control over sentence boundaries and sentence structure; word choice may often be inaccurate. • errors in grammar, spelling, and punctuation interfere with understanding in much of the response.

Score	GRE scoring guide	TOEFL	SAT	NAEP
		• an accumulation of errors in sentence structure and/or usage	• contains errors in grammar, usage, and mechanics so serious that meaning is somewhat obscured	
1	A 1 paper demonstrates fundamental deficiencies in analytical writing. A typical paper in this category exhibits one or more of the following characteristics • provides little or no evidence of the ability to understand and analyze the issue • provides little or no evidence of the ability to develop an organized response • presents severe problems in language and sentence structure that persistently interfere with meaning • contains pervasive errors in grammar, usage, or mechanics that result in incoherence.	An essay at this level is seriously flawed by one or more of the following weaknesses: • serious disorganization or underdevelopment • little or no detail, or irrelevant specifics, or questionable responsiveness to the task • serious and frequent errors in sentence structure or usage	An essay in this category 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	Unsatisfactory response (may be characterized by one or more of the following) • attempts to take a position (addresses topic), but position is very unclear OR takes a position, but provides minimal or no support; may only paraphrase the prompt. • exhibits little or no apparent organization. • minimal or no control over sentence boundaries and sentence structure; word choice may be inaccurate in much or all of the response. • errors in grammar, spelling, and punctuation severely impede understanding across the response.
0	• off-topic (i.e., provides no evidence of an attempt to respond to the assigned topic), in a foreign language, merely copies the topic, consists of only keystroke characters, or is illegible, or nonverbal.	An essay at this level merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.	Essays not written on the essay assignment will receive a score of zero.	Not part of rubric
Not scored	Blank	Not part of rubric	Not part of rubric	Not part of rubric

Appendix B

Glossary of e-rater Microfeatures

Feature	Name of microfeature	Brief description	Example
Grammar	Fragment	A sentence-like string of words that does not contain a tensed verb or that is lacking an independent clause	"And the school too."
Grammar	Run-on sentence	A sentence-like string of words that contains two or more clauses without a conjunction	"Students deserve more respect they are young adults."
Grammar	Garbled sentence	A sentence-like string of words that contains five or more errors, or that has an error-to-word ratio > 0.1, or that is unparseable by the Santa module, which organizes words into clauses	"And except unusual exception, most children can be ease with their parents not the their teachers."
Grammar	Subject-verb agreement	A singular noun with a plural verb or a plural noun with a singular verb	"A uniform represent the school."
Grammar	Ill-formed verb	A mismatch between the tense of a verb and the local syntactic environment; also, use of *of* for *have*, as in *could of*	"We need the freedom to chose what we want to wear."
Grammar	Pronoun error	An objective case pronoun where nominative pronoun is required, or vice versa	"Us students want to express ourselves."
Grammar	Possessive error	A plural noun where a possessive noun should be; usually the result of omitting an apostrophe	"They stayed at my parents house."
Grammar	Wrong or missing word	An ungrammatical sequence of words that is usually the result of a typographical error or of an omission of a word	"The went to their teacher with a complaint."
Grammar	Proofread this!	An error which is difficult to analyze; often the result of multiple, adjacent errors	"They had many wrong science knowladge."
Usage	Wrong Article (Method 1)	A singular determiner with a plural noun or a plural determiner with a singular noun; use of *an* instead of *a*, or vice versa	"I wrote in these book. He ate a orange."
Usage	Articles (wrong, missing, extraneous)	Use of *a* when *the* is required, or vice versa	We had **the good time at the party. (Wrong article) I think it is good for me to share **room with others. (Missing article) I think that mostly people succeed because of **the hard work . (Extraneous article)

Feature	Name of microfeature	Brief description	Example
Usage	Articles (wrong, missing, extraneous)	An article where none should be used or a missing article where one is required	We had **the good time at the party. (Wrong article) I think it is good for me to share **room with others. (Missing article) I think that mostly people succeed because of
Usage	Confused words	Confusion of homophones, words that sound alike or nearly alike	Those young soldiers had to **loose their innocence and grow up. (lose) **Its your chance to show them that you are an independent person. (It's) Parents should give **there children curfews. (their) I think that mostly people succeed because of
Usage	Wrong word form	A verb used in place of a noun	"The choose is not an easy one."
Usage	Faulty comparison	Use of *more* with a comparative adjective or *most* with a superlative adjective	"This is a more better solution."
Usage	Preposition error	Use of incorrect preposition, omitting a preposition, or using an extraneous one	Their knowledge **on physics were very important. (of) The teenager was driving **in a high speed when he approached the curve. (at) Thank you for your consideration **to this matter
Usage	Nonstandard verb or word form	Nonword: Various nonwords commonly used in oral language.	Nonwords: gonna, kinda, dont, cant, gotta, wont, sorta, shoulda, woulda, oughtta, wanna, hafta
Mechanics	Spelling	A group of letters not conforming to known orthographic pattern	
Mechanics	Failure to capitalize proper noun	Compares words to lists of pronouns that should be capitalized (e.g., names of countries, capital cities, male & female proper nouns, and religious holidays)	
Mechanics	Initial caps	Missing initial capital letter in a sentence	
Mechanics	Missing question mark	An unpunctuated interrogative	
Mechanics	Missing final punctuation	A sentence lacking a period	

Feature	Name of microfeature	Brief description	Example
Mechanics	Missing comma or apostrophe	Detects missing commas or apostrophes	Apostrophe: arent, cant, couldnt, didnt, doesnt, dont, hadnt, hasnt, havent, im, isnt, ive, shouldnt, someones, somebodys, wasnt, werent, wont, wouldnt, youre, thats, theyre, theyve, theres, todays, whats, wifes, lifes, anybodys, anyones,
Mechanics	Hyphen error	Missing hyphen in number constructions, certain noun compounds, and modifying expressions preceding a noun	"He fell into a three foot hole. They slipped past the otherwise engaged sentinel."
Mechanics	Fused word	Fused: An error consisting of two words merged together	"It means alot to me." Fused: alot, dresscode, eachother, everytime, otherhand, highschool, notime, infact, inorder, phonecall, schoollife, somethings, noone
Mechanics	Compound word	Detects errors consisting of two words that should be one.	
Mechanics	Duplicate	Two adjacent identical words or two articles, pronouns, modals, etc.	"I want to to go… They tried to help us them."
Style	Repetition of words	Excessive repetition of words	
Style	Inappropriate word or phrase	Inappropriate words. Various expletives.	
Style	And,and,and	Too many sentences beginning with coordinate conjunction	
Style	Too many short sentences	More than four short sentences, less than 7 words	
Style	Too many long sentences	More than four long sentences, more than 55 words	
Style	Passive voice	By-passives: the number of times there occur sentences containing BE + past participle verb form, followed somewhere later in the sentence by the word *by*.	"The sandwich was eaten by the girl."
Development	Number of discourse elements	Provides a measure of development, as a function of the number of discourse elements	
Development	Content development	Provides a measure of average length of discourse elements	

Feature	Name of microfeature	Brief description	Example
Prompt-specific vocabulary usage	Score-group of essays to which target essay is most closely related.	compares* essay to essay-groups 6, 5, 4, etc., and assigns score closest relationship (max cosine). *Cosine of weighted frequency vectors.	
Prompt-specific vocabulary usage	Similarity of essay's vocabulary to vocabulary of essays with score 6	compares* essay to essay-group score 6. *Cosine of weighted frequency vectors.	
Lexical complexity	Sophistication of word choice	Calculates median average word frequency, based on Lexile corpus	
Lexical complexity	Word length	The mean average number of characters within words	