Listening. Learning. Leading.®

Research Report
ETS RR–11-16

Automated Scoring Within a Developmental, Cognitive Model of Writing Proficiency

Paul Deane

Thomas Quinlan

Irene Kostin

April 2011

Automated Scoring Within a Developmental, Cognitive Model of Writing Proficiency

Paul Deane, Thomas Quinlan, and Irene Kostin

ETS, Princeton, New Jersey

April 2011

Technical Review Editor: Jim Carlson

Technical Reviewers: Isaac Bejar and Derrick Higgins

Abstract

ETS has recently instituted the Cognitively Based Assessments of, for, and as Learning (CBAL) research initiative to create a new generation of assessment designed from the ground up to enhance learning. It is intended as a general approach, covering multiple subject areas including reading, writing, and math. This paper is concerned with the writing assessment being developed within the CBAL framework, and in particular, with the potential for using automated scoring techniques effectively within such an assessment to support learning.

Key words: writing assessment, CBAL, automated scoring

Table of Contents

Page

List of Tables

1. Overview

1.1. Context and Purpose

In practice, assessment design involves complex tradeoffs. For instance, many testing programs assess writing directly by administering generic prompts under strict time limits. This type of assessment has gained in popularity relative to indirect measures of writing skill in response to the concern that writing should be assessed by actually requiring students to write, but it has been criticized as encouraging simplistic, formulaic approaches to instruction that fail to develop the complex skills students need to learn in order to become effective writers (Hillocks, 1987, 2002). Portfolio-based writing assessment may encourage richer, more appropriate forms of instruction, but it raises issues of reliability and validity. Automated scoring methods have been introduced and advanced as instructional aids, but they have been criticized as focusing attention too much on mechanical correction of errors rather than encouraging critical engagement with content. In other words, it is no simple matter to design summative assessments that are truly tests worth teaching to (Shepard, 2002). Success in creating tests that truly encourage learning will require a complex, sustained effort that takes into account an extraordinary range of elements, including at least the following:

- the cognitive and instructional literatures that define what students learn and how they best learn it
- psychometric constraints on effective assessment
- institutional and public policy constraints that ultimately determine test feasibility

ETS has recently instituted a research initiative, Cognitively Based Assessments of, for, and as Learning (CBAL), intended to address this problem and create a new generation of assessment designed from the ground up to enhance learning. It is intended as a general approach, covering multiple subject areas including reading, writing, and math (see Bennett & Gitomer, 2009). This paper is concerned with the writing assessment being developed within the CBAL framework, and in particular, with the potential for using automated scoring techniques effectively within such an assessment to support learning. The role of automated scoring for CBAL cannot be discussed, however, without an in-depth presentation of the CBAL initiative. This presentation occupies section 1.2, after which section 1.3 examines the role of automated scoring within this philosophy. The remaining sections of the paper will present studies motivated by the approach sketched in section 1.3, which focuses on the connection between the features used in automated scoring and the development of formative hypotheses that drive instruction.

1.2 Cognitively Based Assessment of, for, and as Learning: The CBAL Initiative

Cognitively based assessment design. The CBAL initiative builds upon the principles of evidence-centered design (Mislevy, Steinberg, & Almond, 2003), which posit that test design should start with an explicit model of what students know and can do (the *student model* or *competency model*) directly linked to evidence about student performance. This linkage is enforced by an *evidence model,* which explicitly identifies what evidence particular classes of tasks will provide about particular competencies. Evidence-centered design provides a philosophy that focuses on working forward from construct to task, first developing an explicit model of competencies to measure, then identifying ways to measure them, and only then designing tasks to fit. In the case of the CBAL initiative, the competency model is explicitly grounded in the cognitive literature.

Drawing upon the literature in writing cognition, Deane, Quinlan, Odendahl, Welsh, and Bivens-Tatum (2008) identified three fundamentally different kinds of skills that should be included in the writing construct:

> Strand I: Expressive language and literacy skills. This is the ability to produce conventional, written English text in an academic or spoken style, as appropriate.
>
> Strand II: Strategies for planning, structuring, evaluating and improving documents. This is the ability to produce well-structured, elaborated discourse, including the process writing skills that support thoughtful construction of extended texts.
>
> Strand III: Critical thinking for writing. This is the ability to reason critically about content, audience, and purpose, and to make use of that reasoning to develop and structure the thought expressed in a text.

These three strands form the basis for a more complex competency model, shown in Figure 1 below.

A full account of this competency model and an explication of the various skills indicated by it can be found in Deane et al., (2008). Certain critical points should be noted. First, it should

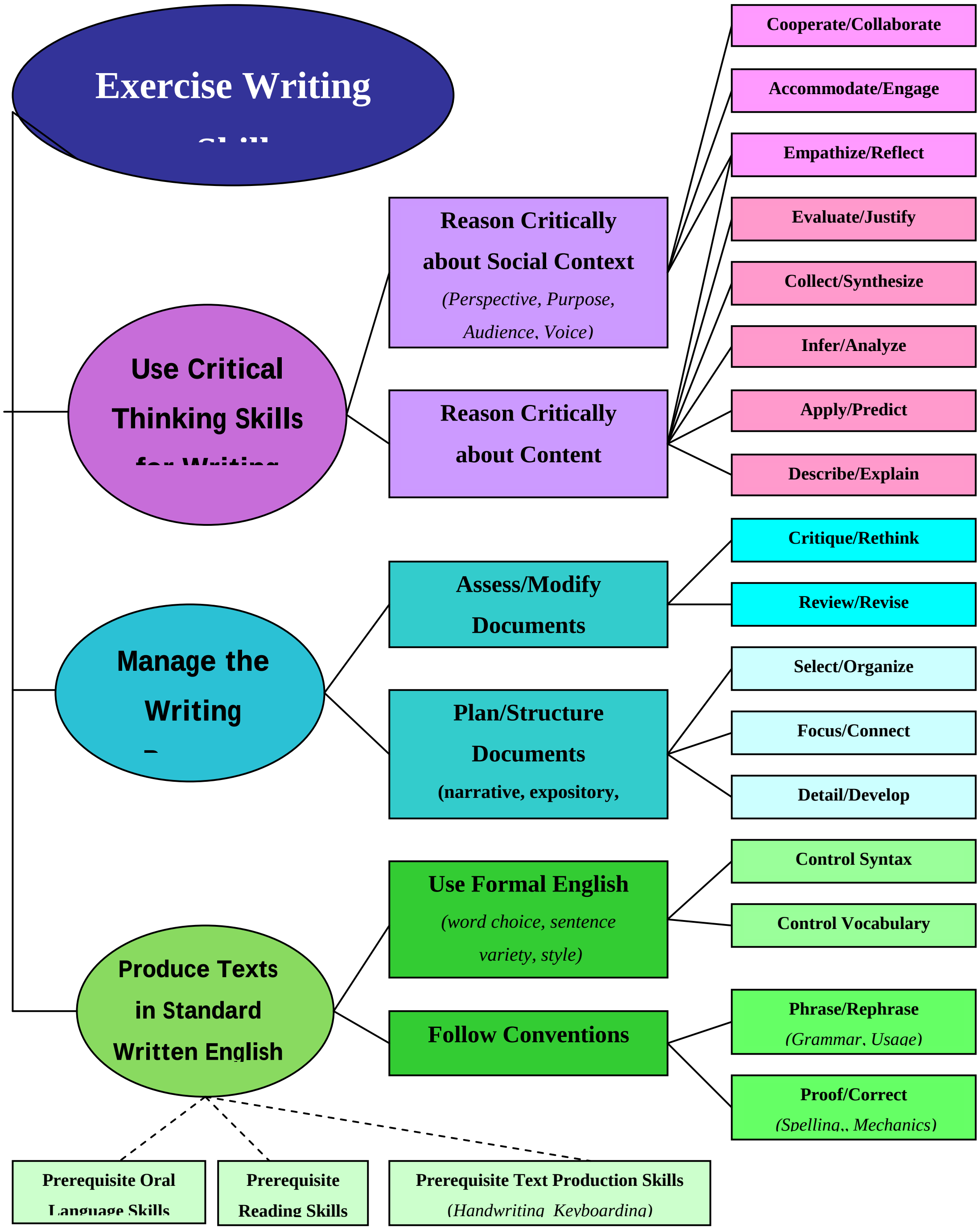

Figure 1. **The Cognitively Based Assessments of, for, and as Learning (CBAL) writing competency model.**

be kept in mind that this model is intended for accountability assessment purposes. It is designed to characterize the competencies required for effective writing and is intended neither as a general cognitive theory nor as a cognitive theory of writing, though it closely reflects the literature on writing cognition.[1] Second, keep in mind that the evidence model for CBAL writing posits a series of evidence rules specifying what features of tasks provide evidence for specific abilities in the competency model. Some are obvious; for instance, spelling mistakes directly provide evidence of (failures in) the ability to *transpose* language into (correctly expressed) text. Others are more abstract, such as the features of a text that instantiate effective argumentation. This link between evidence and competencies means that the model, though expressed abstractly in the diagram, is interpreted very concretely in practice in terms of the features that provide evidence for the skills represented by each node, whether provided automatically by natural language processing or identified as traits by human scorers. These features may be very concrete for those literacy skills represented in the first strand of the competency model, and they may be very abstract for those represented in the third strand; but in either case, a key feature of the model is an attempt to build an explicit evidentiary argument relating observed student performance to the competency model. Tasks selected for the CBAL writing assessment are driven by the requirement that they provide evidence for the full range of constructs specified in the competency model.[2]

Note, critically, that one of the major conclusions built into this competency model is that no easy separation exists among writing skill, critical thinking, and reading skills, and general literacy. At least from a formative perspective, reading, writing, and the acquisition of critical thinking skills are closely related. See Deane et al. (2008) and O’Reilly et al. (2008) for more detailed discussions of these points.

Assessment of, for, and as learning. The purpose of assessment is to encourage learning. Ideally, therefore, an assessment should not only provide good measurement of student achievement (assessment *of* learning), but it should also facilitate future instruction (assessment *for* learning) and, if possible, be a worthwhile educational experience in its own right (assessment *as* learning).[3] The CBAL initiative seeks to design tests that fulfill all of these goals, and in so doing, it responds to common themes in the instructional literature.

For instance, Wiliam (2007) outlined five key strategies that characterize effective teaching across a variety of subject-matters:

- *Sharing learning expectations.* The teacher shares exactly what he or she expects students to learn, not only providing clear expectations but giving them clear criteria for success.

- *Questioning.* The teacher facilitates effective classroom discussions, asks insightful questions, and assigns learning tasks that elicit evidence of student learning.
- *Feedback.* The teacher uses evidence gathered to provide feedback that gives students a clear idea how to move forward.
- *Self-assessment.* The teacher encourages students to take active control of their own learning.
- *Peer assessment.* Students are activated as educational resources for one another.

The strategies outlined in Wiliam (2007) were presented at a high level of abstraction, though they define features we have built into the design for the CBAL summative and formative systems. If we examine reviews of effective instruction more focused on specific content areas, we find similar themes, but with important elaborations and differences. For instance, Langer (2001) identified six characteristics shared by effective middle and high school language arts programs:

- *Skills and knowledge are taught in multiple types of lessons.* Langer (2001) distinguished among separated, simulated, and integrated lessons. *Separated* lessons involve the direct teaching of discrete component skills. *Simulated* lessons require students to exercise these skills in a larger textual frame chosen primarily to support instruction in the target skill. *Integrated* lessons require students to apply these skills in a meaningful context as part of a large, purposeful, and usually complex activity requiring the coordination of many different skills. Langer found that the best classrooms practiced all three types of instruction and systematically coordinated them so that students were expected to apply what they learned in separated or simulated instruction in the larger, more integrated tasks.
- *Tests are deconstructed to inform curriculum and construction.* Langer (2001) distinguished between *separated* and *integrated* test preparation. Separated test preparation focuses on test practice and test-taking strategies. Integrated test preparation focuses on the skills and competencies being tested, and carefully reformulates and restructures lessons to provide students with the skills they need to do well on the test. Langer found that the best classrooms avoided separated test

preparation and embedded requisite skills deeply into the curriculum. This aspect of Langer's model is closely connected with the emphasis on self-assessment in Wiliam (2007).

- *Teachers make connections across instruction, curriculum, and life.* Langer (2001) observed that teachers could point out many different sorts of connections to their students. They could point out connections within lessons, across lessons, across classes, and across grades. They could relate class material to students' lives, both in school and out. Langer found that the best schools consistently use a variety of different ways of forging connections, whereas those schools that performed more poorly made fewer connections or none at all.
- *Students learn strategies for doing the work.* Langer (2001) observed that teachers in the more successful schools consistently provided students with explicit instruction in effective strategies that enable students to break down complex tasks into manageable chunks. Moreover, the effective teachers did more than provide students with strategies showing them how to do a task; they provided strategies that showed them how to think about it. This aspect of Langer's framework is connected to, though hardly identical with, Wiliam's (2007) emphasis on the importance of sharing learning expectations and providing feedback.
- *Students are expected to be generative thinkers.* Langer (2001) observed that teachers in the more successful schools were never satisfied with students learning facts and skills. They probed for connections and implications, expecting students to go beyond the basal level to generate deeper interpretations of texts and produce more richly elaborated ideas in their writing. This aspect of Langer's framework corresponds reasonably closely to Wiliam's (2007) emphasis on teachers using *questioning* as a key pedagogic technique.
- *Classrooms foster cognitive collaborations.* Langer (2001) observed that teachers in the more successful schools structured classroom interactions so that the classroom became a collaborative discourse in which students naturally worked together to deepen their understanding of the material, generate richer interpretations of text, and

sharpen and elaborate their own ideas. This aspect of Langer's outline corresponds reasonably closely to Wiliam's (2007) emphasis on peer assessment.

The correspondence between Wiliam (2007) and Langer (2001) is not exact, but the two frameworks indicate important elements of successful teaching that are clearly applicable to writing instruction.

Finally, Graham and Perin (2007) indicated 11 *best practices* in writing instruction, most of which can easily be interpreted as instantiating the broader frameworks presented by Wiliam (2007) and Langer (2001). They are listed below, reordered in terms of Langer's categories:

- Multiple types of lessons
    - *Sentence combining*, which usually takes the form of focused or simulated lessons that teach students how to build up complex, flexibly organized sentences
    - *Word processing,* which makes use of computers and word processors to support writing instruction
    - These particular techniques are only a small portion of the variety of lessons that would be needed to teach all the skills students need to write well. For any particular writing task, such as writing a research paper, a variety of skills need to be taught—such as paraphrase and summary—that require separated, simulated, and integrated instruction if students are to master them effectively.
- Tests deconstructed
    - *Specific product goals*, which means that students are given clear, specific, achievable goals or rubrics and are taught to evaluate text in terms of those goals
    - *Study of models,* which means that students are given models of good writing and given opportunities to read, analyze, and emulate them
    - The emphasis in Graham and Perin's (2007) presentation suggested that this is a key element of effective writing instruction, and one that requires careful, explicit instruction.
- Connections made

    - *Prewriting,* which means that students are engaged in scaffolded activities that support the generation and organization of content

    - *Inquiry activities,* which means that students are provided with rich content that helps them to develop the ideas presented in their writing

    - *Writing for content learning,* which means that writing is used to help students learn content material and is not kept separate from the rest of the curriculum

- Strategies taught

    - *Writing strategies*, which means teaching students how to manage the complex planning, drafting, revising, and editing processes needed to produce an effective document

    - *Summarization,* which means that students are thoroughly instructed in the strategies required to be able to summarize texts in their own words

- Collaboration fostered

    - *Collaborative writing,* which arranges class work so that students work together as they plan, draft, revise, and edit their texts

    - *Process writing approach,* which creates a workshop environment in which students can work on extended writing projects for real audiences with enough time for multiple drafts, feedback, and revision

Put together, these kinds of results suggest specific ways in which tests could be structured so that they (a) function better as learning experiences in their own right, (b) fit more naturally into classrooms that employ best practices, and (c) encourage best practices by modeling them. These goals correspond with many of the emphases adopted by the CBAL initiative; in fact, the connection to Langer's (2001) model is particularly strong:

- Multiple types of lessons: CBAL tests use a mixture of preliminary (separated or simulated tasks) and richer, integrated tasks.

- Tests deconstructed: The tasks used in CBAL tests are selected so that they correspond to tasks teachers would naturally use to teach the competencies being assessed, creating a natural bridge between test preparation with instruction. Instead of a single year-end assessment, multiple tests are administered during the school year, creating a natural linkage between assessment and instruction, and making it possible for assessment results to affect instruction while it is still in progress. In addition, multiple assessments during the school year make it possible to cover the construct more fully, assessing a wider range of competencies through an array of foundational tasks covering common writing purposes and genres. Multiple tests also encourage integration of the assessment into the instructional sequence.

- Connections made: The tasks in CBAL tests are connected and set in a realistic school or real life context, forming a natural project-like sequence that makes explicit connections across tasks and relates the test both to the classroom and real life.
- Strategies taught: The tasks in CBAL tests are *scaffolded*, that is, they implicitly model a strategy for solving a particular type of complex, integrated task, and each step is provided with appropriate supporting materials that make the intended strategy explicit.
- Generative thinking encouraged: CBAL tests emphasize constructed response tasks that require students to generate thoughtful answers.
- Collaboration fostered: Many CBAL tasks are designed to fit naturally into a collaborative context, even if the test itself does not directly deploy collaborative tasks.

A complex task, consisting of a set of coordinated tasks that displays these characteristics, is referred to in the CBAL context as a *foundational task set*. This term is intended within the CBAL context specifically as a technical term to describe sequences of tasks that have the features listed above (Deane, Fowles, Baldwin, & Persky, 2011; O'Reilly et al., 2009). In CBAL test design, foundational task sets have an important role both in the construction of summative assessments (since normally, a single CBAL assessment would consist of a tasks instantiating key parts of a single foundational task set) and in formative assessments designed to support classroom instruction directly. In the CBAL writing assessments, the foundational task sets are intended to instantiate best practices in writing instruction, with the sequence of tasks that may appear on a summative assessment particularly focusing on prewriting, inquiry activities, writing for content learning, and effective use of writing strategies, while also providing appropriate rubrics and models.[4]

ETS is exploring ways to incorporate as many of the best practices in writing instruction into foundational task sets as possible and developing formative and summative materials based upon them. At this point in the process the development of these task sets, and building in appropriate connections with instruction and pedagogy, is very much a work in progress, and we expect to revise and develop the ideas in depth as we proceed. But the kinds of goals we have set for ourselves can usefully be explicated by considering one example in depth. Let us therefore consider a task set developed for formative use, in particular, to support persuasive essay writing, which critically exercises the evaluate/justify node in the CBAL competency model. In a set of tasks provided to teachers in the 2007 CBAL pilot, the following project was introduced:

> Imagine that at your school everyone is discussing whether or not junk food (unhealthful food and drinks) should be sold at the school. You and your classmates are trying to learn more and make up your own minds. In this project, you can research the issue, explore arguments on both sides of the issue, and write an essay for your school newspaper to explain your point of view.

Ten tasks were presented, as follows:

Part 1. Evaluating and Choosing Different Sources

Task 1. Evaluate types of information and sources.

Task 2. Use guidelines to evaluate an Internet source.

Task 3. Make a T-chart of arguments for and against inviting a speaker.

Task 4. Argue for choosing one speaker and against choosing other speakers.

Part 2. Building Your Own Argument

Task 5. Consider arguments for selling junk food in school.

Task 6. Consider arguments against selling junk food in school.

Task 7. Present your view in an essay.

Task 8. Consider ways to revise for a different audience.

Part 3. Reviewing Someone Else's Argument

Task 9. Consider ways to improve an introduction.

Task 10. Explain the strong and weak points of an argument.

These tasks are presented in a coordinated set that includes reading matter, rubrics/guidelines, and other supporting materials.

A summative version of this task set would necessarily include a subset of tasks in order to fit within reasonable time constraints; conversely, teachers' instructional plans might expand considerably upon this outline. But the set of 10 tasks presented here illustrate in rough, preliminary form what is meant by a foundational task set. We hope to develop a series of such task sets, motivating each in terms of both the competency model and in terms of the literature on best instructional practice.[5]

1.3. Automated Scoring for Cognitively Based Assessments of, for, and as Learning (CBAL) Writing

In 2007, ETS conducted small pilots of draft CBAL writing assessments. One of the goals of this work was to experiment with methods that would enable CBAL writing assessments to be scored with respect to performance on the CBAL writing competency model. In the scoring plan for these pilots, holistic scoring was avoided in favor of a partially analytic scoring system that distinguished among the three strands of the CBAL competency model. That is, human ratings were assigned with respect to Strand I (expressive language and literacy skills), Strand II (strategies for planning, structuring, evaluating, and improving documents), and Strand III (critical thinking skills for writing). The scoring methods applied in this pilot were intended to explore ways in which CBAL foundational tasks could be used to provide evidence for skills specified in the CBAL competency model.

While the pilots were entirely human scored, the CBAL initiative intends to use automated scoring as much as possible consistent with its other goals (e.g., to support learning and effective instruction). Ideally, CBAL tests should be scored almost immediately, with test results being returned to students and teachers to support self-activated learning, indicate what further formative assessments are appropriate, and facilitate effective instructional intervention. This goal, combined with the desire to make heavy use of constructed-response formats, strongly indicates the need to use automated essay scoring (AES).[6]

AES scoring has a long history within writing assessment, dating back to the seminal work of Page (1966). More recently, several approaches to AES have come into use for scoring large-scale assessments, including the PEG system, a descendant of Page's original system (Page, 1994; 1995), methods based on Latent Semantic Analysis (LSA; Foltz, Laham, &

Landauer, 1999; Landauer, Laham, & Foltz, 2000), and the *e-rater*® system (Attali & Burstein, 2006; Burstein et al., 1998). The key question is the extent to which AES can be applied effectively in the CBAL context, where there is a strong need for reliable scores on multiple dimensions—and an accompanying need for useful formative feedback—without compromising constructs where automated scoring cannot capture the essential features.

Historical background. AES can be viewed as a special case, an application of a research tradition that uses features of the text to measure properties of cognitive and linguistic interest, such as essay quality, text readability, genre characteristics, or mastery of language. One branch of this tradition can be found in the literature on second language development in writing. Wolf-Quintero et al. (1998) provided an excellent summary of this literature by discussing existing measures focused primarily on measurements of fluency, lexical and syntactic complexity, and accuracy in writing using such features as clause length and T-unit length, the normalized number of word types, the ratio of subordinate clauses to clauses, and the percent of text without error. These kinds of features have been shown to provide useful measures of the development of language skill among second language learners. This particular line of research, while focused within second language learning, reflects an earlier literature that examined various proxies for fluency and syntactic complexity, such as the T-unit (essentially, a main clause with any subordinate clauses attached to it). Hunt (1970), for instance, demonstrated that T-units increase in length as students mature, with shorter T-unit lengths in early elementary school and longer T-unit lengths in college and adult writing. However, as Crowhurst (1983) demonstrated, no strong, direct relation exists between T-unit length and writing quality, either as a direct predictor or as a result of intervention designed to increase T-unit length. Similarly, other features of both spoken and written language are known to increase with age and maturity for native speakers, roughly paralleling the literature for foreign language learners, including vocabulary richness and sentence complexity (Loban 1976).

The entire line of AES research focuses on proxies for fluency, complexity, and accuracy in language production that ultimately relates back to the literature on readability (Dale & Chall, 1948; Dale & Tyler, 1934; Flesch, 1948; Lively & Pressey, 1923; Ojemann, 1934; Patty & Painter, 1931; Thorndike, 1921; Vogel & Washburne, 1928) in which proxies for length and vocabulary difficulty were used to predict differences in text reading difficulty. More recent research tradition addresses related issues. This is the quantitative approach to genre analysis

(Biber, 1988, 1995a, 1995b; Biber, Conrad, & Reppen, 1998). In this approach, a wide range of features are extracted, instead of just a few, and their behavior over a large corpus is investigated by performing a factor analysis. The factors represent dimensions of variation, such that different genres, or types of text, represent characteristic 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 writing that reflect the maturation and development of the writer and corresponding variations in texts that make them easier, or harder, for readers to understand; these seem in turn to correspond to the factors that underlie variation among texts.

Automated scoring and related work at ETS. The primary AES in use at ETS is the e-rater system, which has been driven in large part by continuing efforts to make the system more transparent with respect to constructs with which writing instructors are concerned. While the original e-rater system used more than 50 individual features and selected whichever features best predicted human scores for a particular prompt, e-rater 2.0 uses a much smaller set of features specifically selected to measure particular aspects of writing performance, such as content, organization, development, vocabulary, grammar, usage, mechanics, and style (Burstein & Shermis, 2003; Burstein, Chodorow, & Leacock, 2004; Attali & Burstein 2006; see also Ben-Simon & Bennett, (2007).)[7]

Attali and Powers (2008) took steps to ground e-rater in empirical developmental data and use it to build a developmental writing scale applicable to 4th through 12th grades. They reported the results of a large empirical study in which a national sample of 12,000 students, drawn from 170 schools and more than 500 classes from 4th, 6th, 8th, 10th, and 12th grade, wrote essays to four prompts each. The essay assignments were distributed in a block design in which adjacent grades shared essay topics. Student essays were collected online and scored using e-rater 2.0.

Their key findings were (a) it was possible to develop a single writing scale that uses e-rater features to score student essays from 4th to 12th grades and (b) the best such analysis involved three underlying factors: fluency (as measured by essay length and the e-rater style feature), conventions (as measured by the e-rater grammar, usage, and mechanics features), and word choice (as measured by the e-rater vocabulary and word length features). As Attali and Powers (2008) noted, however, the scale is “based only on the specific features used in this study, and thus is limited by what those features measure. Even though past research showed very high correlations between e-rater scores and human scores, it is clear that important dimensions of writing are not represented in this feature set" (Attali & Powers, 2008, p. 57).

Another method of automated text classification at ETS is SourceFinder™, which includes a module that classifies texts by genre and grade level using techniques inspired by Biber's genre analysis (Deane, Sheehan, Sabatine, Futagi, & Kostin, 2006; Sheehan, Kostin, & Futagi, 2007a, 2007b). In particular, Sheehan et al. (2007a) examined a corpus of potential source documents for use by ETS testing programs, where the intent was to filter texts so that they correctly instantiated genres of interest to test developers.[8] Six factors were extracted: spoken language, academic discourse, overt expression of persuasion, oppositional reasoning, sentence complexity, and unfamiliar vocabulary. Deane et al. (2006) examined variations in a set of 3rd through 6th grade texts drawn from a large corpus of materials typically used in school and included features characteristic both of Biber's genre analysis and readability studies. 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.[9] Sheehan et al. (2007b) 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 the following five factors: academic style, sentence complexity, vocabulary difficulty, subordination, and oppositional reasoning. These five factors were used to develop predictive grade-level models (i.e., separate models for expository and literary texts) in which academic style, subordination, and oppositional reasoning contributed to the prediction of grade level alongside the sentence complexity and vocabulary difficulty factors more typically used in readability measures. In addition, many of these features, including the spoken language factor in particular, were used in models predicting expository versus literary genre.

E-rater and SourceFinder represent relatively independent lines of work and make use of fairly distinct feature sets. Together, they capture most of the types of automatically identified features that have been proposed for measuring syntactic maturity, readability, and essay quality. Large datasets are available at ETS for both feature sets, including the Attali and Powers (2008) corpus, which provides a national writing sample across multiple grades. What both strands of research have in common is a sustained effort to identify features that correspond to appropriate writing and literacy constructs rather than using isolated features selected for predictive value.

Applying automated scoring within the Cognitively Based Assessments of, for, and as Learning (CBAL) Framework. The CBAL initiative places summative assessment within a formative context, which means, ultimately, a developmental context. As a result, student performance may need to be evaluated in multiple ways and for multiple purposes. For example, the CBAL writing assessment pilots discussed at the beginning of this section are specifically pilots of a summative test in which automated features would be used to score actual performance on test items in order to make high stakes decisions within an accountability assessment. This is one of the classic contexts for AES. Another, very different context is provided by the ETS product Criterion^{TM}, a product that scores actual performance to provide feedback to students as part of classroom instruction. Given the design priorities of the CBAL initiative, we can expect other purposes for testing to be relevant. The following distinctions are in order:

- *Nature of prediction.* Are automated features being used for the following:
    - To score actual performance on test items?
    - To predict future performance (on grades or other test scores)?
    - To provide normative or developmental information?
- *Nature of the stakes.* Will automated features be used for the following:
    - To make high-stakes decisions about schools, teachers, or individuals?
    - To inform classroom teaching by suggesting formative hypotheses to teachers?
    - To provide feedback to students as part of instruction?

At least the following combinations of prediction, stake, and context are likely to be relevant from a CBAL perspective:

1 Scoring items on the high-stakes accountability assessments administered periodically as part of the CBAL writing assessment (*summative scoring*).

2 Using (earlier) accountability assessments to predict end-of-year aggregate scores in order to inform classroom teaching or even using automated features to predict human scores on the same test, providing much earlier score information than otherwise possible (*early warning*).

3 Using accountability assessments to provide teachers with normative information that suggests a range of formative hypotheses to be investigated in greater detail (*preliminary indication*).

4 Using diagnostic tests or activities to provide teachers with normative or developmental information that confirms or disconfirms particular formative hypotheses (*diagnosis*)

5 Scoring actual performance on classroom tasks in such a way as to provide feedback to students while helping suggest formative hypotheses to teachers (*formative scoring*)

Each of these possible applications implies somewhat different constraints on the use of automated scoring features, though in each case, it is important to be clear about the construct these features are measuring. For high-takes scoring, predictive value is critical, and features will be selected for the efficacy with which they can separate levels of performance. For formative purposes, features might be selected because they provide evidence for particular interpretations of student performance.

Construct representation. Regardless of the purpose for which automated scoring (NLP) features are selected to use as the basis for a scoring model, it is critical that they have clear connections to appropriate writing constructs. Ideally, we would like to know not only whether features are appropriate, but also how they relate to other features that might be considered for scoring or as the basis for a formative hypothesis, and it would also be valuable to have a clear picture how these features develop over time. To achieve these goals, we will consider a set of factor analyses that were constructed based upon the Attali and Powers (2008) developmental dataset. These factor analyses suggest several underlying factors that can easily be related to reasonable writing constructs. Section 2 will present this factor analysis in detail and will analyze how the factors identified map onto the CBAL competency model.

Summative scoring and e-rater. Of existing e-rater features, most of them (word length, word frequency, grammar, mechanics, usage, and style) can be interpreted as providing evidence of the first strand in the CBAL competency model (fundamental language and literacy skills; Attali & Powers, 2008). But two (organization and development) can be interpreted as providing evidence for the second strand in the CBAL competency model, insofar as they measure whether students have produced texts containing internal subdivisions, each of which is developed to some reasonable length. It is thus an open question whether current e-rater features could be redeployed, possibly with some additional features drawn from the SourceFinder feature set, to provide direct prediction of Strand I and Strand II scores. It is very probable and arguably appropriate that Strand III scores, focusing on critical thinking, will require human not automated scoring. Section 3 of this paper will therefore focus on potential methods of predicting human scores for the different strands of the CBAL competency model, but it will also explore implications with respect to using the accountability assessments to generate early warnings indicating whether students are on track to meet standards by the end of the year.

Feature selection for formative purposes. While a summative assessment is not a diagnostic assessment, CBAL periodic assessments should provide enough information about individual performance to suggest formative hypotheses that teachers may wish to follow up on. For example, a student obtaining a lower-half score on the CBAL writing assessment might display a combination of features typical for their score point (little organization and development, multiple grammatical and mechanical errors, relatively simple vocabulary), and that combination might co-occur with other features not used for scoring, such as little syntactic variety or the overuse of typical spoken elements such as the first person pronoun. Depending on the exact configuration of features, various hypotheses about students' skill levels might be suggested, which teachers could then use to inform their instruction. In the example given above, for instance, one might wonder whether student performance reflects a lack of high level writing strategies (yielding essays without significant organization or development), whether it reflects a failure to adopt an academic style, whether it reflects a lack of some specific skill (such as paraphrasing, spelling or vocabulary), or whether it reflects all of these in combination. It is unlikely that the summative assessments will provide enough information to disambiguate among hypotheses, but even at this low level of precision, such information is likely to prove instructionally useful. Moreover, since the final CBAL assessment system is likely to include automatically scored diagnostic assessments and classroom exercises, it is important to identify features that provide useful normative or developmental information and thus suggest important formative hypotheses, even if they are not used in a summative scoring model. Section 4 of this paper will therefore review a range of features, both those directly implicated in scoring and others that correlate with known, relevant factors, and will elaborate a range of hypotheses about how some of these features could be used for scoring or feedback.

2. Modeling Writing Development and Writing Constructs: An Approach Through Factor Analysis

The CBAL writing competency model is in its design a model of expert performance and is thus focused on the end-state (the desired complex of knowledge and abilities) and not on the developmental process by which maturing students increase their writing skill. Presumably, however, increased maturity as writers and increased skill at writing are linked, in which case one would expect that strong connections would exist between the features that govern text quality and the features that change as writers mature. Similarly, connections are likely between reading and writing, since writers are unlikely to produce texts much more complex than the texts that they are able to read. These considerations argue that much can be learned by constructing a Biber-style factor analysis of writer's texts, parallel to (but not necessarily the same as) the kinds of factor analyses that have been developed for collections of edited documents.

Viewed in this light, much can be learned from the factor analyses developed for the ETS SourceFinder application by Sheehan and associates. Sheehan et al. (2007b) examined a corpus of texts intended for readers ranging from early primary to high school and developed a factor analysis that yielded nine factors which included the following five factors: academic style, sentence complexity, vocabulary difficulty, subordination, and oppositional reasoning. These five factors were used to develop predictive grade-level models (i.e., separate models for expository and literary texts) in which academic style, subordination, and oppositional reasoning contributed to the prediction of grade level alongside the sentence complexity and vocabulary difficulty factors more typically used in readability measures.

The dimensions obtained in the most recent factor analysis in this line of research (Sheehan et al., 2010) involve the features and dimensions shown in Table 1.

To our knowledge no comparable analysis has been performed with a corpus of student writing, but the technique has obvious potential for defining dimensions of text variation across a corpus that illustrates student writing development. The Attali and Powers (2007) corpus, introduced above, has precisely the desired properties since it covers student writing in a national sample across multiple grades, and therefore it supports a factor analysis in which developmental

Table 1

Dimensions Underlying Variation Among Texts in a Corpus of School-Appropriate Texts

Feature	Loading
Dimension 1: Spoken vs. written language	
First person singular pronouns	+.98
First person plural pronouns	+.96
Communication verbs (ask, call, question, etc.)	+.74
Wh words (who, what, where, etc.)	+.62
Contractions (didn't, can't, I've, etc.)	+.60
Conversation Verbs (get, know, put, etc.)	+.60
Mental state verbs (appreciate, care, feel, etc.)	+.54
Question marks	+.53
Attributive adjectives	–.46
Non–proper nouns	–.81
Dimension 2: Academic orientation	
Nominalizations (–tion, –ment, –ness, –ity)	+.90
Academic words (Coxhead)	+.81
Abstract nouns (existence, progress, etc.)	+.77
Cognitive process/perception nouns	+.65
Academic verbs (apply, develop, indicate, etc.)	+.63
Average characters per word (log chars.)	+.61
Clarification conjuncts (for example, namely)	+.46
Passive constructions	+.33
Average concreteness rating (MRC database)	–.75
Dimension 3: Narrative style	
Past tense verbs	+.79
Past perfect aspect verbs	+.78
Third person singular pronouns (he, she, etc.)	+.61
Present tense verbs	–.86
Dimension 4: Sentence complexity	
Average no. of subordinate clauses	+.99
per sentence	+.95
Average no. of words per clause (log words)	+.92
Average no. of words per sentence (log words)	+.38
Prepositions	
Dimension 5: Vocabulary difficulty	
TASA SFI < 30 (token count)	+.89
TASA SFI < 30 (type count)	+.81
Average TASA SFI	–.64
Dimension 6: Overt expression of persuasion	
To infinitives	+.85
Necessity modals (should, must, etc.)	+.64
Conditional subordinators (if, unless, etc.)	+.56
Possibility modals (can, can't, could, etc.)	+.52
Predictive modals (will, would, etc.)	+.51
Dimension 7: Negation	
Synthetic negation (no, neither, nor)	+.71
Adversative negation (alternatively, etc.)	+.63
Negative adverbs (never, seldom, rarely)	+.55

Note. From *Generating automated text complexity classifications that are aligned with targeted text complexity standards* by K. M. Sheehan, I. Kostin, Y. Futagi, and M. Flor, 2010, ETS Research Report No. RR-10-28, Princeton, NJ: ETS. Copyright 2010 by Educational Testing Service.

features are relevant. There is thus reason to believe that similar techniques could be applied to build a developmental model of student writing.

The feature sets underlying SourceFinder and Criterion provide, between them, an extensive set of linguistic features measuring many aspects of English vocabulary, grammar, meaning, and style. It is not clear, a priori, however, whether the way these features are aggregated for either of these products provides the best measurement for the purpose of measuring Strand I of the CBAL writing competency model. We therefore undertook a systematic reanalysis of these features as applied to the large collection of essays that formed the basis for Attali and Powers' (2007) developmental writing scale. We extracted both feature sets, calculated them for the essays, and then performed a factor analysis comparable to those undertaken in previous investigations (Deane et al., 2006; Sheehan … Futagi et al., 2006; Sheehan et al., 2007a). We then examined ways to use these factors, or to aggregate features in other ways, in order to provide measurement of writing performance that simultaneously provided good measurement while being readily interpretable in light of the CBAL competency model.

2.1 Initial Analysis

Data preparation. The Attali and Powers (2008) dataset was divided into four subsets based upon essay order, that is, the point in the school year that each essay was taken. (Each student wrote four essays, staggered across the school year.) Since some participants were lost as the Attali and 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 only one essay per participant: either a persuasive essay or a descriptive one. Each essay was processed using ETS's natural language processing software to produce a dataset containing a large set of candidate features for use in later analysis. The automatically calculated features incorporated in this analysis included not only features that have been selected for use in the operational Criterion and SourceFinder systems, but also certain component features that have always been aggregated operationally with other features in the e-rater scoring engine, plus a range of features that have not been operationally deployed but which belong to feature sets developed for research purposes.[10]

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.

Exploratory factor analysis. Exploratory factor analysis was performed upon each of the essay orders, and multiple factor analyses were considered in order to obtain the most interpretable set of factors across essay orders. The analyses reported below reflect an initial exploratory analysis using Principal Components Analysis, followed by a factor analysis using principle axis factoring and Promax rotation. A common analysis with 10 factors replicated across all four essay orders. Table 2 presents the results. Several points are worth noting.

First, two of the factors are partial replications of aggregations of features used in e-rater. The verb error factor draws entirely from features used to calculate the e-rater grammar feature. Similarly, the orthographic accuracy factor draws primarily from features used to calculate the e-rater mechanics feature, though one feature (confusion of homonyms) is attracted to this factor rather than to other features from the usage category to which it is assigned in e-rater. On the other hand, some features, such as the wrong, missing, or extraneous article feature from e-rater, may be problematic because they load positively on the noun-centered text factor, which could indicate that it is measuring the presence of nouns more than it is measuring the presence of article errors.

Second, several of the factors are strongly allied with those obtained in earlier work (e.g., factors having to do with academic orientation, sentence complexity, spoken language/personal stance, persuasion, and narrative style). These factors employ many of the same features that have repeatedly surfaced in earlier analyses of written, edited text.

These results suggest that student writing displays many of the same dimensions of variation as edited, published text and partially confirm some of the feature aggregations used in e-rater, while raising questions about specific features.

2.2 Second Order Factor Analysis

A second order factor analysis was performed on 7 of the 10 factors identified in the exploratory analysis. The two genre factors were excluded (narrative style, overt expression of persuasion), as well as the tenth factor, which only attracted a single feature. Table 3 shows the pattern matrices that resulted over the four essay orders.

Table 2

Exploratory Factor Analyses of the Attali/Powers (Attali & Powers, 2008) Dataset

Feature	Essay order 1	Essay order 2	Essay order 3	Essay order 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)				
Abstract nouns (existence, progress, etc.)	+.76	+.74	+.79	+.75
Passive verbs	+.69	+.57	+.66	+.63
Median word frequency	+.52	+.52	+.46	+.43
Dale List of Common Words	–.63	–.60	–.60	–.58
Imageability score (MRC database)	–.84	–.85	–.82	–.86
	–.85	–.85	–.81	–.88
Dimension 2: Noun-centered text				
Definite determiners (log per 1,000 words)	+.92	+.89	+.98	+.87
Wrong, missing or extraneous articles	+.68	+.72	+.70	+.74
Noun/verb ratio	+.56	+.60	+.60	+.56
Nouns (log per 1,000 words)	+.52	+.56	+.55	+.53
Document length in words	+.35	+.38	+.34	+.38
Dimension 3: Sentence complexity				
Verbs (log per 1,000 words)	+.96	+.92	+.98	+.94
Average sentence length in words	+.93	+.89	+.95	+.92
Prepositions (log per 1,000 words)	+.48	+.49	+.48	+.58
Dimension 4: Spoken style				
Mental state verbs (appreciate, care, feel, etc.)				
Conversation verbs (get, know, put, etc.)	+.78	+.84	+.78	+.79
First person singular pronouns				
Noun/verb ratio	+.67	+.68	+.74	+.74
Attributive adjectives	+.30	+.37	+.41	+.45
(log per 1,000 words)	–.36	–.33	–.33	–.35
	–.75	–.58	–.73	–.75
Dimension 5: Overt expression of persuasion				
Predictive modals (will, would, etc.)				
Conditional subordinators (if, unless, etc.)	+.84	+.86	+.87	+.86
Present tense	+.74	+.75	+.80	+.71
	–.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
Present tense verbs	–.52	–.43	–.55	–.40
Dimension 8: Orthographic accuracy				
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

Table 3

Second Order Factor Analyses Over the Four Essay Orders on 7 of the 10 Factors Identified by the Exploratory Factor Analysis

	Component 1				Component 2				Component 3			
	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	-.21	-.34	-.26	-.32	**.74**	**.68**	**.72**	**.67**	.03	.02	-.02	.10
Orthographic accuracy	-.14	-.07	-.06	-.02	-.26	-.36	-.32	**-.85**	**.79**	**.61**	**.69**	.27
Verb errors	-.12	-.06	.04	-.06	.33	.26	.25	-.10	**.74**	**.87**	**.86**	**.93**

Note. Values representing strong feature loadings are in bold.

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 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. This observation can be confirmed by examining the cross-correlation between e-rater features and factor scores, as shown in tables B1 through B4 in Appendix B.[11] In particular, these tables indicate obvious associations between e-rater features and the three macrofactors that we have identified. Thus, 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 second 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 varied in alignment in one essay order, lining up with component 2 (fluency) instead of component 3 (accuracy),[12] the overall trend is very clear. The

Attali and Powers (2008) developmental data clearly reflects three components when a wide range of features are included.

2.3 Discussion

An examination of the connection between e-rater features and the factor structure indicates that an exact correspondence does not exist; some e-rater features appear to mix information drawn from more than one factor. Thus, the style feature correlates strongly with fluency and with the spoken-to-academic dimension, and the grammar feature correlates almost as well with sentence complexity, elaboration, and academic orientation as it does with the verbal error factor. Since these and other features are aggregated by combining a large collection of more specific error detection features, it is possible that these features reflect differential behavior of some of their component features. The behavior of e-rater features with respect to factor scores is presented in tables A5 through A15 in Appendix A. An analysis of these results suggests that some cause for concern may exist with respect to the e-rater style feature for the following reasons:

1 Two of the features (too many long sentences and agentive passives) have positive correlations with human score, counter to expectations for what is essentially an error feature. Since the human scores for the Attali and Powers (2008) data are not intended to be used for validation, not too much weight should be placed upon this fact, but it may be worth examining in other datasets to see whether the observed trend is stable.

2 Some of the features (too many long sentences and too many short sentences) are most strongly connected with the sentence complexity factor, but others (repetition of words and agentive passives) are most strongly connected negatively or positively with the academic-to-spoken factors (academic orientation, noun-centered text, spoken style).

3 At least for this dataset, the repetition of words feature appears to be doing most of the work, and it correlates most strongly with the spoken style factor. Correlations for other features are small or in the wrong direction. It might therefore make sense to pull this feature out on its own, possibly in combination with other features that weigh on the spoken-to-academic language dimension.

These three points suggest that the style feature may inappropriately aggregate features connected to sentence complexity with features connected with the academic-to-spoken language dimension, though we cannot say so conclusively on the basis of this dataset alone. One of the usage features (confusion of homonyms) also shows ambiguities, correlating with both grammatical and spelling error features, and with the corresponding factors (orthographic accuracy and verb errors). In addition, three features (hyphenation errors and two article error features) correlate strongly with the noun-centered text factor and show weakly positive correlations with human scores in this data, again raising questions that would need to be examined with a larger dataset. All of the features except spelling errors and repetition of words are very sparse, so it is dangerous to read too much into the correlations observed, though they are certainly suggestive.

The interesting question, however, is how these factors (and the corresponding e-rater features) can be interpreted in light of the CBAL writing competency model. A natural connection is found at the second level of the competency model. The features going with the first second order factor (i.e., those associated with academic orientation, noun-centered text, and negatively with spoken style) can naturally be interpreted as measuring student mastery of academic English and development of the ability to control when and whether to use a more oral style, and thus can be associated with the master academic English node. Similarly, the third second order factor, being associated with e-rater error features for grammar and mechanics, naturally corresponds to the follow written conventions node.[13]

The elaboration and sentence complexity factors, however, may be more closely related with Strand II, though an argument can also be mounted simply relating them to document length and hence to fluency in text production. Features associated with the elaboration factor in particular include various connective elements arguably present in sustained, well-developed text, the presence of paragraph breaks, and other features suggesting complex, well-developed documents. Thus, the e-rater organization and development features correlate strongly with the elaboration factor, as do various other SourceFinder features reflecting text elaboration, including the type-token ratio and measurements of the number and length of paragraphs, as shown in Table 4.

Table 4

Strong Correlates of the Elaboration Factor

Feature	Correlation with elaboration factor
Log of document length in words	.655
Document length in words	.627
Indefinite pronouns (someone, anyone, no one, etc.)	.536
Wh-determiners (which, whose)	.467
Adversative conjunctions (yet, but, etc.)	.451
e-rater organization feature	.405
Length of longest paragraph in words	.382
Average length of paragraph in words	.368
e-rater development feature	.260
Wh-adverb (when, why)	.260
Average sentence length	.250
Concessive subordinator (although, etc.)	.247
Number of paragraphs	.229
Coordinating conjunctions	.226
…	…
Noun-verb ratio	-.345
Rate of word repetition	-.401
Sentences per 1,000 words	-.523

What these features reflect, viewed in combination, is the ability to produce sustained texts with clear internal structure, reflected in paragraph breaks, the use of transitional words of various sorts, and the presence of sustained development of content, as expressed in the e-rater development feature and the use of longer paragraphs. Absence of this ability is reflected in shorter texts, shorter paragraphs, and shorter sentences, with relatively repetitive use of vocabulary. Thus, while this factor is closely correlated with the document length feature, it is arguably not mere fluency; rather, it reflects fluency in producing complex, structured texts with clear internal subdivisions and development of ideas at the paragraph level. This is precisely what Strand II of the CBAL competency model is concerned with. The strong connection with document length is natural, given the impossibility of producing structured texts without also producing longer texts, and need not be viewed as necessarily problematic as long as the features used to measure this construct are sensitive to the presence of structure in the document and not to length alone.

While the human scores available with the Attali and Powers (2008) dataset should not be relied on too closely, since they were collected for very different purposes under unusual scoring conditions,[14] a clear competition exists between the academic orientation factor and the e-rater features we have associated with Strand I. In particular, we observed that if we combine standard e-rater features with the academic orientation factor and build a regression model against human score, the academic orientation factor takes the place of the word length and style features for descriptive essays, though without an overall improvement in performance.[15]

It should be noted that Strand II as specified in the competency model includes not only the ability to produce sustained, structured texts, but also the ability to draw upon a complex assemblage of writing strategies useful in producing such texts. The features in the Attali and Powers (2008) dataset provide no direct measurement of this aspect of the model, though it is conceivable that measures of writing behavior, rather than the written product, might provide additional measurement. This possibility will be examined further below.

Put together, the results reported in this section are consistent with two major conclusions:

- First, that the features associated with e-rater appear to measure Strands I and II of the CBAL competency model, though not Strand III, which is very strongly concerned with the quality of thought.
- Second, that Strands I and II can also be measured by a broad range of additional features, associated with the primary factors in the factor analysis, that are not part of e-rater but which correlate with one of the key dimensions identified in the factor analysis.

3. Some Initial Results From Cognitively Based Assessments of, for, and as Learning (CBAL) Pilot Tests

As mentioned in Section 1.3, two CBAL pilot writing assessments were administered in Fall 2007. The pilots were given to relatively small populations in three middle schools in an urban/suburban school district. The first form (referred to below as Form K, based upon its initial numbering in a set of pilot designs) focused on persuasive writing. It had a pro/con organization, in which students were expected to understand and deal appropriately with arguments for or against a school policy (whether junk food should be sold on school campuses). The second form (referred to below as Form M) also focused on persuasive writing but dealt primarily with the evaluation of alternatives (specifically, which of three class projects should be adopted for a service learning project). The tests were administered using a spiraled design, so that each test sampled a different random subgroup drawn from the same base population.

The primary purpose of the pilot was to try out the tasks deployed in each test design and to test strand-specific scoring techniques in which each writing prompt was scored separately for Strand I (language and literacy skills), Strand II (strategies for producing documents), and Strand III (critical thinking for writing). As part of this trial, several types of features were collected to test the potential for automated scoring: e-rater features and micro-features, features drawn from SourceFinder and allied research, and features based upon keystroke logs, which provide an indication of pause lengths during text production.

Three constructed-response writing tasks were included on each test: two short-answer (paragraph-length) questions and one requiring an essay-length response. Each response was scored manually by human raters using strand-specific rubrics prepared for the purpose. Results on both forms indicated reasonably high correlations between tasks but also across strand scores within a task. On Form K, correlations within the same strand across tasks ranged from .45 to .69, with the majority of correlations above .60. [Kim: Should these be two digits?] Correlations across strands and tasks ranged from .38 to .63, with the majority of correlations above .50. On Form M, correlations within the same strand across tasks ranged from .47 to .88, with the majority of correlations above .65. Correlations across strands and tasks ranged from .41 to .83, with the majority of correlations above .70, although correlations between the long and short tasks were largely in the range between .50 and .70. Reliability as measured by Cronbach's alpha was above .90 on both forms for the correlation between individual tasks and total test score, and about .80 for the correlations across tasks within strands and across strands within tasks.[16] Exploratory factor analysis suggested a single underlying construct, rather than clearly separating the strand scores as factors. However, we investigated whether it would be possible to use regression analysis to predict some of these traits from the associated features despite their close correlations. Some initial results looked promising in this regard, but the sample size was too small for definitive conclusions. The overall reliability of the test appeared to be high and to measure a unified writing construct, though the samples were too small to permit detailed conclusions to be drawn.

3.1 E-rater Features and Strand Scores

Several stepwise regression analyses were constructed using e-rater features to predict individual and total scores by strand, focusing on predicting scores for the long essay responses. These analyses were consistent with the hypothesis that most e-rater features measure Strand I of the CBAL competency model, with the organization and development features accounting for much of the prediction of Strand II scores.[17]

Regression analyses by strand. For Form K, stepwise regression yielded the following models (Tables 5 and 6) when e-rater features alone were entered into the analysis.

Table 5

Best Stepwise Regression Model for Essay Response, Form K, Predicting Strand I Essay Scores With e-rater Features Alone

Feature	Standard coefficients (beta)	Significance
Mechanics	.40	<.001
Grammar	.33	.002
Word length	.29	.001
Usage	.27	.002
Style	.19	.01

Note. *R* = .92, *R*-square = .85, adjusted *R*-square = .83.

Table 6

Best Stepwise Regression Model for Essay Response, Form K, Predicting Strand II Essay Scores With e-rater Features Alone

Feature	Standard coefficients (beta)	Significance
Organization	.73	<.001
Development	.33	.001
Style	.31	.001
Usage	.21	.025

Note. *R* = .89, *R*-square = .80, adjusted *R*-square = .77.

For Form M, fewer features made it into each model (see Tables 7 and 8), but they were the strongest (and distinctive) features in the Form K models; that is, grammar and mechanics (for Strand I) versus organization and development (for Strand II).

Table 7

Best Stepwise Regression Model for Essay Response, Form M, Predicting Strand I Essay Scores With e-rater Features Alone

Feature	Standard coefficients (beta)	Significance
Grammar	.54	.001
Mechanics	.51	.001

Note. *R* = .86, *R*-square = .73, adjusted *R*-square = .70.

Table 8

Regression Model for Essay Response, Form M, Predicting Strand II Essay Scores

Feature	Standard coefficients (beta)	Significance
Organization	1.059	<.001
Development	.349	.003

Note. *R* = .93, *R*-square = .86, adjusted *R*-square = .85.

These partialed-out models compare favorably in their overall accuracy to standard e-rater models and are consistent with the hypothesis that the e-rater organization and development features measure aspects of Strand II whereas the grammar and mechanics features measure aspects of Strand I in the CBAL competency model.

It is important to note at this point that no similar pattern exists with respect to Strand III scores assessing critical thinking for writing. For Form K, the best model involves a mixture of most e-rater features; and for Form M, the Strand III model primarily depends on the same features as Strand II, as shown in Tables 9 and 10. These results are consistent with the hypothesis that none of the e-rater features directly measures critical thinking for writing, and they predict Strand III scores indirectly through Strand I and II correlates.[18]

3.2 Extending the Model With SourceFinder Features

The factor analysis presented in Section 2 was intended in part to suggest ways in which the CBAL automated scoring model could be enriched. In particular, that analysis suggested that the e-rater scoring model focuses on a subset of the factors that it identifies: elaboration (through the organization and development features), academic orientation (in the word length and word frequency features), orthographic accuracy (through the mechanics feature), and verb errors (primarily through the grammar feature). However, at least for the Attali and Powers (2008) data, automated scoring with factor scores did not improve on the base e-rater model. But the existence of the factor structure identified over the Attali and Powers data suggests the possibility that scoring could be improved by including additional features, particularly features that load on dimensions not currently exploited by the e-rater scoring engine. An examination of extended scoring models for the Attali and Powers data suggested that such features may be useful, a conclusion that appears to extend to the CBAL pilot data, where use of features drawn from the SourceFinder feature set yields improved models.

Table 9

Best Stepwise Regression Model for Essay Response, Form K, Predicting Strand III Essay Scores With e-rater Features Alone

Feature	Standard coefficients (beta)	Significance
Organization	.60	<.001
Style	.36	<.001
Development	.33	.001
Grammar	.27	.005
Word length	.23	.008

Note. *R* = .89, *R*-square = .79, adjusted *R*-square = .76.

Table 10

Best Stepwise Regression Model for Essay Response, Form M, Predicting Strand III Essay Scores With e-rater Features Alone

Feature	Standard coefficients (beta)	Significance
Organization	.87	<.001
Development	.34	.02
Style	.28	.03

Note. *R* = .92, *R*-square = .84, adjusted *R*-square = .81.

The regression models discussed below are intended as purely exploratory analyses and should be read in that light. Our goal is to identify whether any features appear to add variance in the particular datasets we considered and not to prove definitively that these features should consistently be included in a scoring model. A fuller analysis would also use factor scores and examine their role as features. Such an analysis, for the Attali and Powers (2008) data did not perform as well as the existing e-rater features, taken as a whole, but at least one factor (the academic orientation factor) did appear to provide some variance. Given the nature of regression analyses, it is quite possible that an analysis using multiple component features from a factor will perform better than an analysis using factor scores directly.

In particular, we should note that when features associated with additional dimensions were included for human-scored persuasive essays from the Attali and Powers (2008) dataset, features associated with sentence complexity (proportion of verbs in sentence) and academic orientation (abstract nouns, academic verbs) added .01 to the model's *R*-square, as shown in Tables 11 and 12.

Table 11

Stepwise Regression Model Against Human-Scored Persuasive Essays From the Attali and Powers (2008) Dataset

Feature	Standard coefficients (beta)	Significance
Organization	.69	<.001
Development	.43	<.001
Word length	.16	<.001
Grammar	.14	<.001
Style	.10	.001
Mechanics	.07	.003
Median word frequency	-.06	.006

Note. *R* = .85, *R*-square = .72, adjusted *R*-square = .71.

The major difference between the two models is that the two new academic-dimension features replaced median word frequency, which also loads on the academic orientation factor, and that sentence length also comes into play through the proportion of verbs in sentence feature (where a high proportion of verbs corresponds normally to very simple sentences).

A similar pattern emerges for the descriptive essays from the Attali and Powers (2008) dataset (see Tables 13 and 14). The baseline model is improved, though only marginally, by additional features reflecting elaboration (loading on average paragraph length), word frequency and academic verbs (loading on academic orientation), and by proportion of verbs per sentence (loading negatively on sentence complexity).

Table 12

Stepwise Regression Model Using a Combination of SourceFinder and e-rater Features: Persuasive Essays From the Attali and Powers (2008) Dataset

Feature	Standard coefficients (beta)	Significance
Organization	.67	<.001
Development	.43	<.001
Word length	.17	<.001
Grammar	.13	<.001
Style	.11	<.001
Academic verbs	.08	.001
Proportion of verbs in sentence	-.08	.001
Mechanics	.06	.006
Abstract nouns	.06	.007

Note. *R* = .86, *R*-square = .73, adjusted *R*-square = .72.

Table 13

Stepwise Regression Model Against Human-Scored Descriptive Essays From the Attali and Powers (2008) Dataset

Feature	Standard coefficients (beta)	Significance
Organization	.67	<.001
Development	.45	<.001
Word length	.18	<.001
Mechanics	.13	<.001
Grammar	.09	.008
Usage	.08	.006
Style	.07	.04

Note. *R* = .84, *R*-square = .70, adjusted *R*-square = .70.

Table 14

Stepwise Regression Model Using a Combination of SourceFinder and e-rater Features: Descriptive Essays From the Attali and Powers (2008) Dataset

Feature	Standard coefficients (beta)	Significance
Organization	.73	<.001
Development	.35	<.001
Average paragraph length	.17	.012
Mechanics	.12	<.001
Word length	.11	.002
Grammar	.09	.008
Mean word frequency (lexile corpus)	-.07	.035
Usage	.07	.019
Academic verbs	.07	.008
Proportion of verbs in sentence	-.06	.021

Note. *R* = .84, *R*-square = .71, adjusted *R*-square = .70.

Similar results can be observed for prediction of CBAL pilot test strand scores. As Table 15 illustrates, additional variance can be accounted for by adding features loading on relevant factors. In the case of Form K, adjusted r-square increases from .76 to .80 for Strand I and from .76 to .77 for Strand II with the addition of a few key features. In the case of Form M, adjusted r-square increases from .81 to .93 with added features for Strand I and from .86 to .89 for Strand II. The relevant models are show in Tables 15 through 18.

The features that add prediction to these models are appropriate to the strand and thus fit into the general picture we have developed thus far in which Strand I and Strand II human scores are predicted by different and construct-appropriate features. For Strand I, the relevant features are concentration of nouns (loading on the noun-centered text factor), academic verbs (loading on the academic orientation factor), and mental state verbs (loading on the spoken style factor), all of them loading on the academic versus spoken dimension in the second order analysis. For Strand II, the relevant features are average paragraph length (competing with the more linguistically sophisticated e-rater development feature) and proportion of verbs in the sentence (loading on the sentence complexity factor, and thus belonging to the same second order factor as the elaboration factor that dominates Strand II).

Table 15

Stepwise Regression Model Using a Combination of SourceFinder and e-rater Features: Cognitively Based Assessments of, for, and as Learning (CBAL) Pilot, Form K, Predicting Strand I Human Scores

Feature	Standard coefficients (beta)	Significance
Mechanics	.56	<.001
Usage	.33	<.001
Style	.28	.001
Concentration of nouns	.23	.006

Note. See also Table 7. *R* = .91, *R*-square = .82, adjusted *R*-square = .80.

Table 16

Stepwise Regression Model Using a Combination of SourceFinder and e-rater Features: Cognitively Based Assessments of, for, and as Learning (CBAL) Pilot, Form K, Predicting Strand II Human Scores

Feature	Standard coefficients (beta)	Significance
Organization	.75	< .001
Average paragraph length	.34	.001
Style	.25	.006
Usage	.24	.01

Note. See also Table 8. *R* = .89, *R*-square = .79, adjusted *R*-square = .77.

Table 17

Stepwise Regression Model Using a Combination of SourceFinder and e-rater Features: Cognitively Based Assessments of, for, and as Learning (CBAL) Pilot, Form M, Predicting Strand I Human Scores

Feature	Standard coefficients (beta)	Significance
Grammar	.64	<.001
Mental state verbs	.49	<.001
Mechanics	.44	<.001
Style	.32	.001
Academic verbs	.18	.019

Note. See also Table 9. *R* = .97, *R*-square = .94, adjusted *R*-square = .93.

Table 18

Stepwise Regression Model Using a Combination of SourceFinder and e-rater Features: Cognitively Based Assessments of, for, and as Learning (CBAL) Pilot, Form M, Predicting Strand II Human Scores

Feature	Standard coefficients (beta)	Significance
Organization	.82	<.001
Development	.41	<.001
Proportion of verbs in sentence	-.23	.04

Note. See also Table 10. *R* = .97, *R*-square = .90, adjusted *R*-square = .89

These results are suggestive, but not conclusive. It seems likely, however, on the basis of these preliminary results, that a more extensive CBAL dataset would make it possible to predict Strand I and Strand II human essay scores at least as accurately as e-rater currently predicts human scores and that the resulting model would do so in a principled way, using construct-appropriate features.

3.3 Extending the Model With Behavioral Features

All of the features considered thus far are product features, measured by examining the final product of student work. However, written products provide only a partial picture of students' writing skill, as posited by the CBAL writing competency model. The act of writing unfolds over time, as the writer produces a text, and this unfolding reveals something about students' underlying cognitive processes. Pauses during text production have been shown to associate strongly with the hierarchical structure of a text, suggesting that pauses reflect the relative difficulty of problem-solving (Matsuhashi, 1981; Schilperoord, 2002). One of the goals of the CBAL pilot was to trial an additional class of feature: process features measuring the behavior of candidates while writing. These features, particularly those connected to production of short spans of text (characters, words, possibly sentences) have the advantage that they reflect the basic skills of text production as they operate in real-time. Thus, they might be relatively useful to create a more direct measurement of the efficiency of individual text production processes. Several features were collected, in particular:

- Length of pauses between characters within a word
- Length of pauses between words
- Length of pauses between sentences[19]
- Time taken to produce single backspaces within a word
- Time taken to produce single backspaces after a word break
- Time taken to produce a sequence of multiple backspaces
- Number of characters in a burst, defined as a sequence of text produced with no pause greater than one second

Means and medians were calculated for exploratory purposes; this is the data presented below.[20] There is already evidence in the literature that keystroke log data of this sort can be related to writing quality (REFS). On a construct basis, we would expect that certain features (e.g., those happening at word boundaries or within a word) to reflect motor processes of text production (e.g., the inscribe and transpose nodes in the CBAL competency model), while others (e.g., pauses between sentences, or long sequences of backspaces) might reflect global, text-planning processes to a greater extent.

On Form K, however, none of these behavioral features showed significant correlations with essay score without taking other variables into account. On Form M, however, three of these features displayed significant correlations with human scores, as shown in Table 19. But when the behavioral features were included in stepwise regression modeling, timing features added variance to the Form K models (see Tables 20 and 21) but did not do so to the (already quite accurate) predictive models associated with Form M. The features that added variance were in fact consistent with the constructs: Longer pauses between words predicted poorer performance on Strand I, whereas more time spent backspacing predicted higher performance on Strand II.[21] Since Form K is the form with a larger, though still small *N* (40 cases), no definitive conclusions can be drawn, but the results suggest that behavioral features of this type may well provide useful measurement of both Strand I and Strand II competency.

Table 19

Pearson Correlations Between Behavioral Features and Human Strand Scores for Form M

Feature	Strand I	Strand II	Strand III
Mean burst length	.72 *p* < .001	.58 *p* < .01	.52 *p* < .02
Mean pause between characters within a word	-.49 *p* < .03	-.51 *p* < .02	-.48 *p* < .03
Mean pause between words	-.51 *P* < .02	-.51 *p* < .02	-.50 *p* < .02

Table 20

Stepwise Regression Model Using a Combination of SourceFinder, e-rater and Behavioral Features: Cognitively Based Assessments of, for, and as Learning (CBAL) Pilot, Form K, Predicting Strand I Human Scores

Feature	Standard coefficients (beta)	Significance
Mechanics	.56	<.001
Usage	.29	.001
Style	.27	.001
Concentration of nouns	.24	.003
Mean pause between words	-.16	.036

Note. See also Table 15. *R* = .92, *R*-square = .84, adjusted *R*-square = .82.

While the small sample size and tentative nature of the tasks on the two CBAL pilots preclude any broad generalizations without further research, the analyses conducted thus far are highly suggestive. The results suggest that the e-rater feature set may well usefully be supplemented by features that provide better measurement of the academic to spoken dimension, in particular, and by features that provide direct measurement of fluency in text production and time devoted to specific tasks such as backspacing, which in turn implicates revision. Even more importantly, a picture tentatively emerges from these analyses in which strand scoring is meaningful, in that the human-assigned strand scores, though highly correlated, are predicted by different, and construct-relevant, variables.

Table 21

Stepwise Regression Model Using a Combination of SourceFinder, e-rater Features, and Behavioral Features: Cognitively Based Assessments of, for, and as Learning (CBAL) Pilot, Form K, Predicting Strand II Human Scores

Feature	Standard coefficients (beta)	Significance
Organization	.84	<.001
Average paragraph length	.35	<.001
Style	.26	.002
Mean time spent in multiple-backspace sequences	.19	.023
Usage	.18	.037

Note. See also Table 16. *R* = .91, *R*-square = .82, adjusted *R*-square = .80.

It should be noted that in these analyses linear regression has been employed, not because we believe it is necessarily ideal, but primarily for comparability with existing e-rater models and also as a way of exploring possible extensions to the construct implied by those models. It is an open question whether nonlinear models, or models that allow co-prediction by multiple, partially redundant features, might perform better at predicting human scores. Similarly, considerable caution must be exercised in the interpretation of the human-calculated strand scores for the CBAL pilot tests. The results are encouraging, in that they suggest that the strand scores are meaningfully differentiable (at least with respect to Strands I and II). They are also encouraging, in that they suggest that strand scores can be predicted automatically for two of the three strands posited in the CBAL writing competency model. What these results suggest, taken together, is the need for much larger studies, sufficient to support the construction of generalizeable scoring models for Strands I and II, if possible across multiple grades.

4. Going Beyond Summative Single-Test Models

Thus far our discussion has implicitly focused upon one of the uses identified in section 1: summative scoring. We have not explicitly considered other possible applications, in particular, early warning, preliminary indication, diagnosis, or formative scoring.

4.1 Early Warning

We cannot, strictly speaking, determine whether CBAL writing assessments can be used for early warning purposes until multiple assessments have been administered to a single class over the course of a school year, at which point it will be possible to determine how accurately performance on an early test will predict overall performance by the end of the school year. However, some of the regression analyses we performed demonstrated that features of the long essay task could be used quite successfully to predict overall test score and thus to provide early warning in the form of preliminary, predictive (but not final) scores in advance of human scoring of some part of the test.

For Form K, the best regression analysis obtained was that shown in Table 22.

Table 22

Prediction of Total Test Score From Features of the Long Essay Task (Form K)

Feature	Standard coefficients (beta)	Significance
Organization	.442	< .001
Median burst length	.347	< .001
Usage	.294	< .001
Cohesion[a]	.257	< .001
Style	.250	< .001
Mechanics	.182	.005
Vocabulary concreteness score	-.151	.013
Average paragraph length	.150	.05

Note. R = .96, R-square = .92, adjusted R-square = .90.

[a] The cohesion feature used here measures overlap of word stems between each sentence and the previous two sentences. This feature correlates .31 with the academic orientation factor in the Attali and Powers (2008) data.

For Form M, with its relatively small *N*, the best regression analysis obtained was that shown in Table 23.

While, as before, these results should be viewed with extreme caution, they indicate that overall test score can be predicted quite reliably from features of just one (albeit the longest, most reliable) task. Extension from predicting test scores to predicting overall performance for the entire school year may or may not be feasible, but these results are at least suggestive. Note, however, that both of these models make critical use of a behavioral feature (burst length or pause length between words), both of which can be viewed as a measure of the efficacy of automatic text production processes. While such features might be questioned for the purpose of summative scoring—where the quality of the final product may reasonably be argued to be the only truly valid method of scoring—they can easily be justified for early warning purposes.

Table 23

Prediction of Total Test Score From Features of the Long Essay Task (Form M)

Feature	Standard coefficients (beta)	Significance
Organization	.538	<.001
Mean pause between words	-.436	<.001
Mechanics	.311	.003

Note. *R* = .93, *R*-square = .88, adjusted *R*-square = .86.

4.2 Gathering Evidence About Individual Nodes in the Competency Model

Each of the three remaining functions we have identified for a CBAL assessment (preliminary indication, diagnosis, formative scoring) require having a clear idea of what sort of formative hypotheses teachers might wish to entertain; and this, in turn, is closely related to the competency model. That is, if a student performs poorly on Strand I, that fact by itself does little to suggest intervention strategies to the teacher. But if a student were to display some specific profile of behavior that implicated a specific node in the competency model, the teacher would have a much clearer idea how to proceed. For instance, if a student in fact produced a rich vocabulary, but seemed to have serious trouble with spelling, that might suggest one strategy; whereas a very different strategy might be indicated for a student who seemed incapable of using a rich, written vocabulary yet produced very few misspelled words. These kinds of variations in performance may provide rather distinct student profiles, which may be useful for score reporting or formative assessment purposes, even if they do not form the basis for an overall summative score.

In the discussion up to this point, the mapping from e-rater features to nodes in the competency model, while discussed, has not been explored in depth, in part because accuracy of prediction was the first concern (as long as the features were clearly relevant to the constructs we wished to measure, either for formative or summative purposes). In order to provide information that will support feedback at the classroom level (e.g., to the teacher[22]), the features reported on must be closely tied to the competency model, even if they are not identical to the features used in a scoring model. It is in fact difficult to produce a model using linear regression that outperforms e-rater scoring models, except by introducing document length directly as a variable. However, a regression model forces the choice of one of a set of relatively collinear variables, and in many cases, other variables from such a set (or the general trend they illustrate) may provide more transparent construct representation, at least for formative purposes.

In our case, we have identified a number of factors that correlate with human scoring and which themselves have obvious interpretations in terms of the competency model. For instance, the dominant features in the academic orientation factor correspond to properties of words known to affect vocabulary difficulty: a word's familiarity combined with its phonological and morphological complexity. Most of these features (like those in other factors correlated with document quality) show a strong correlation not only with essay quality, but with student grade level, since writing skill shows a clear, but gradual developmental trend upward, presumably reflecting student acquisition of language and literacy skill.

An obvious strategy suggests itself since we have access to the large Attali and Powers (2008) dataset, in which all essays are classified by grade level, even if they are not assigned human essay quality scores. Given a node in the competency model for which we have known, relevant features, it makes sense to construct a feature designed to measure this feature, using grade level data to train it. Such features will be easy to interpret and thus may be useful to support formative assessment, even if it is not ideal for purposes of assigning overall scores.

In our case, we can plausibly construct (or identify) such features for several of the factors identified in Section 2 and interpret those features in terms of the competency model, as follows in Table 24.

These assignments provide us with a fairly clear picture of what aspects of the competency model are currently covered by automated scoring: all of Strand I except the prerequisite speaking and reading skills; large parts of Strand II (but not the critical/evaluative aspects that underlie high level critiquing, editing and revision), and nothing directly covering Strand III.

Table 24

Connections Between Factors, Features, and Cognitively Based Assessments of, for, and as Learning (CBAL) Competency Model Nodes

Factor or feature	Node(s) in the competency model
Academic vs. spoken (vocabulary features loading on academic orientation)	Master academic English
Academic vs. spoken (syntactic features loading on academic orientation and spoken style dimensions)	Master academic English
Academic vs. spoken (spoken style features)	Master academic English
Sentence complexity (e-rater style feature)	Plan/structure documents[23]
Elaboration	Plan/structure documents
e-rater organization feature	Select/organize
e-rater development feature	Detail/develop
Orthographic accuracy (e-rater mechanics feature)	Follow written conventions
Timing features (word level)?	Prerequisite text production skills follow written conventions
Verb errors (e-rater grammar and usage features)	Follow written conventions[24]

In what follows we will use the Attali and Powers (2008) data to develop features corresponding to several of these factors (written vocabulary, written style, and sentence complexity) and then examine how they play out when applied to the CBAL pilot data. This should be considered entirely exploratory, as considerable follow-up work would need to be done to validate the features and examine their applicability in a formative context. But any attempt to formulate a formative system requires first that such features be available and be interpretable in terms of the competency model.

Use written vocabulary. Use written vocabulary is one of the low-level nodes postulated in the CBAL competency model. The factor analysis suggests a range of features that appear to measure level of vocabulary: frequency features, such as mean lexile word frequency; word lists, such as the Coxhead list of typically academic vocabulary; word length features, such as the number of syllables in a word, and morphological features, such as negative prefixes and nominalizations.

We developed an aggregated feature by building a regression model in which these features were used to predict grade level, using the first and second essay orders as the training set and cross-validating on the third and fourth essay orders. While this aggregated model is optimized to predict a different construct—grade level—it is a useful proxy, given that we know that writing quality tends to increase with grade level. The grade-level-based proxy gives us a way to see how closely this relationship holds and whether it is constant over all essay types. To the extent that the feature's predictive value changes or remains constant, we can gain some insight into how vocabulary growth by grade contributes to essay quality.

The training set produced the following model (see Table 25).

Examining the feature's correlation with grade level and with human holistic scores (where available) demonstrated the following pattern of behavior (see Table 26).

Table 25

Regression Model Predicting Grade Level From Vocabulary Features From the Attali and Powers (2008) Dataset

Feature	Standard coefficients (beta)	Significance
Representation of words from the combined Coxhead academic word list	.217	<.001
Mean lexile word frequency	-.201	<.001
Average number of syllables in a word	.098	<.001
Representation of typically Academic verbs	.095	<.001
Representation of words from a list of abstract nouns	.05	<.001
Representation of words with negative prefixes	.05	<.001
Representation of nominalizations	.05	<.001

Note. R = .56, R-square = .31, adjusted R-square = .31.

Table 26

Correlations With Grade Level and Essay Quality for Written Vocabulary Feature Based Upon the Regression Model in Table 24

Dataset	Correlation with grade level	Correlation with persuasive essay scores	Correlation with descriptive essay scores
Essay orders 1 & 2 (training)	.56 (*N* = 10,090, *p* < .001)	.60 (*N* = 297, *p* < .001)	.38 (*N* = 295, *p* < .001)
Essay order 3 (cross-validation)	.53 (*N* = 4162, *p* < .001)	.44 (*N* = 176, *p* < .001)	.43 (*N* = 105, *p* < .001)
Essay order 4 (cross-validation)	.49 (*N* = 3284, *p* < .001)	.23 (*N* = 116, *p* < .02)	.65 (*N* = 168, *p* < .001)

The results are reasonable, though the variation in the cross-validation sets should be noted. It is probably not safe to assume that all prompts induce the same use of vocabulary; indeed, there are probably prompts where there is a strong motivation for writers to avoid complex, academic vocabulary. However, the stability of the prediction of grade level across training and cross-validation sets is encouraging.[25]

This feature assigns values that range approximately between 6 and 12, corresponding roughly to predicted (a slightly curtailed) grade level scale. On Form K, the mean for this feature is 8.99 and the standard deviation is 1. On Form M, the mean for this feature is 9.2 and the standard deviation is .82. Table 27 shows how this feature performs on the two CBAL pilot prompts.[26]

The essay Strand I scores and the total Strand I scores (shown in bold) show the strongest and most significant correlations, consistent with the hypothesis that this feature is primarily measuring a component skill for Strand I. Moreover, the features combined in this written vocabulary feature represent important aspects of the construct, since academic vocabulary tends to be more abstract, more phonologically, and morphologically complex, and less frequent than everyday spoken vocabulary. It would be interesting in future work to examine in detail whether a feature like this one, based upon multiple construct properties, would correspond better to intuitions about student growth in vocabulary knowledge.

Table 27

Performance of Written Vocabulary Feature on Cognitively Based Assessments of, for, and as Learning (CBAL) Pilot Prompts

Criterion	Form K (N = 39)	Form M (N = 22)
Strand I long essay score	.50 (p < .001)	.57 (p < .01)
Strand II long essay score	.23 (p = .16)	.44 (p < .05)
Strand III long essay score	.30 (p = .06)	.52 (p < .02)
Total Strand I score	.54 (p < .001)	.69 (p < .001)
Total Strand II score	.29 (p = .074)	.41 (p < .07)
Total Strand III score	.34 (p < .04)	.52 (p < .02)
Total score	.41 (p = .01)	.56 (p < .01)

Use written style. Use written style is another low-level node in the competency model, focusing on what is often termed *syntactic variety* in the pedagogical literature, although properly speaking it involves mastery of the stylistic patterns typical of written language, and thus the ability both to use typical written/academic syntactic patterns and to avoid spoken features in written contexts where they are less appropriate. One complexity of this competency model node is that it covers more than just the spoken style factor; a variety of syntactic features from the academic versus spoken metafactor are relevant. In order to construct a feature that models this node of the competency model using all of the features that appear to be relevant (e.g., that load in the right directions on the relevant factors and predict grade level and human essay scores), we constructed multiple regression analyses over the same training data to predict grade level and created a feature that represented the mean between two equally predictive sets of features drawn from the factors loading on the academic versus Spoken second order factor. The resulting equation is shown in Table 28 (displaying raw weights, not normalized loadings).[27]

Examining this feature's correlation with grade level and human scores revealed the pattern shown in Table 29. There were very strong correlations with human scores across the board, without a clear preference for one strand over another, despite the theoretical assignment

Table 28

Weights for a Written Style Feature

Feature	Coefficient in predictive equation
(constant)	–0.21
Present tense verb forms	0.65
Wh-determiners	0.53
Adjectives	0.51
Definite articles	0.44
Academic downtoners (barely, hardly, etc.)	0.41
Possessive determiners (my, our, your, etc.)	0.41
Perfect aspect verb forms	0.39
Focus adverbs (only, also, etc.)	0.38
Negative universal quantifiers (no, none, etc.)	0.38
Wh-adverbs	0.32
Passives	0.27
Demonstratives (this, these, those)	0.25
To-infinitive verb forms	0.25
Reflexive pronouns (myself, ourselves, etc.)	0.23
Past tense verb forms	0.22
Multiple-function subordinators	0.21
Adversative conjunctions (alternatively, etc.)	0.21
Subset quantifiers (many, some, few)	0.17
Negative adverbials (seldom, rarely)	0.17
Wh-pronouns	0.17
Sentence negation (not, won't, can't, etc.)	0.15
Indefinite articles (an, an)	0.13
Concessive subordinators (although, though)	0.13
Adverbials of time (later, now, etc.)	0.12
Cohesion (overlap of stems with 2 preceding sentences)	0.10
Emotion words (afraid, amuse, etc.)	–0.21
Causal subordinator (because)	–0.23
Conditional subordinators (if, unless)	–0.29
Mental state verbs (appreciate, care, etc.)	–0.30
2nd person pronouns	–.032
3rd person pronouns	–0.32
Communication verbs (ask, call, etc.)	–0.37
1st person plural pronouns	–0.39
Conditional adverbs (maybe, perhaps, etc.)	–0.40
Additive conjunctions	–0.59
Exclamation marks	–1.02

of this feature to Strand I.[28] The feature appears to correlate fairly robustly with human score in the Attali and Powers (2008) dataset and, more critically, to provide even stronger prediction of human score in the CBAL pilot data, as shown in Table 30.

Table 29

Predictive Strength of a Written Style Feature

Dataset	Correlation with grade level	Correlation with persuasive essay scores	Correlation with descriptive essay scores

Essay Orders 1 & 2	0.55	0.56	0.28
(training)	(*N* = 10,090, *p* < .001)	(*N* =297, *p* < .001)	(*N* =295, *p* < .001)
Essay Order 3	0.52	0.40	0.38
(cross-validation)	(*N* = 4,162, *p* < .001)	(*N* =176, *p* < .001)	(*N* =103, *p* < .001)
Essay Order 4	0.47	0.22	0.56
(cross-validation)	(*N* = 3,284, *p* < .001)	(*N* =116, *p* = .018)	(*N* =168, *p* < .001)

Table 30

Prediction of Cognitively Based Assessments of, for, and as Learning (CBAL) Pilot Scores by the Written Style Feature

Criterion		
	Form K (N = 39)	*Form M (N = 22)*
Strand I long essay score	.56	.60
	(*p* < .001)	(*p* = .003)
Strand II long essay score	.51	.60
	(*p* = .001)	(*p* = .003)
Strand III long essay score	.46	.62
	(*p* = .003)	(*p* = .002)
Total Strand I score	.47	.56
	(*p* = .002)	(*p* = .007)
Total Strand II score	.52	.66
	(*p* = .001)	(*p* = .001)
Total Strand III score	.45	.68
	(*p* = .003)	(*p* = .002)
Total score	.52	.68
	(*p* = .001)	(*p* = .001)

The interpretation of this feature is very straightforward in terms of the underlying construct, since the positive pole corresponds, roughly, to the notion of syntactic variety, with the positively weighted features representing the choice of a rich variety of referential devices for nouns and tense choices for verbs, the use of discourse connectives to handle focus shifts and contrasts, and a higher than usual incidence of various marked syntactic constructions such as passives and reflexives. The negative pole is interpreted straightforwardly as marking use of conversational style, with its preference for a relatively small set of connectives (such as the word *because*, the primary causal subordinator), heavy use of pronominal reference, and even heavier use of syntactic devices marking personal viewpoint and perspective. It would be very interesting in future work to examine whether this feature corresponds reasonably well to teacher intuitions specifically about academic versus spoken style, and whether it can provide a useful measurement of students' progress over time.

Sentence complexity. In the discussion above, we suggested that sentence complexity should be viewed as a (partial) measure of the plan/structure document competency model node, to be used in combination with the written vocabulary and written style features. We built a fairly simple model of sentence complexity over the Attali and Powers (2008) dataset, using grade level as the training criterion. The following regression model resulted (see Table 31).

Table 31

Regression Model Predicting Grade Level From Sentence Complexity

Feature	Standard coefficients (beta)	Significance
Sentences per 100 words	-.24	p < .001
Concentration of prepositions	.24	p < .001
Concentration of verbs	-.32	p < .001
Average sentence length	.37	p < .001

Note. R = .52, R-square = .27, adjusted R-square = .27.

Examining this feature’s correlation with grade level and human scores revealed the pattern of correlations shown in Table 32, with similar performance on CBAL pilot scoring as shown in Table 33.

Note that when the written vocabulary feature is substituted for the two vocabulary features in regression models for this data, it performs within about .01 r-squared of the existing e-rater features. The other features we have discussed (written style, sentence complexity) generally do not make it into the regression models due to overlaps with e-rater features. Thus it is not clear whether a written style feature or a sentence complexity feature would help summative scoring, but there is a clear value in providing measurement of these features (which are highly predictive of grade level and in fact are much stronger predictors than e-rater features even on the cross-validation sets). A combination of these three features with standard e-rater features is able to predict grade level very effectively. That is, in all four essay orders in the Attali and Powers (2008) data, a regression on grade level returns these three features as the strongest predictors, with a multiple R of about .72 and an r-square of about .51. The usefulness of these features in predicting grade level means that they might be very useful for formative purposes—as a measurement of student progress—even if they were not the most strongly predictive features for summative scoring.

Table 32

Correlation of the Sentence Complexity With Grade Level and Human Scores From the Attali and Powers (2008) Dataset

Dataset	Correlation with grade level	Correlation with persuasive essay scores	Correlation with descriptive essay scores
Essay orders 1 & 2 (training)	.55 (*N* = 10,090, *p* < .001)	.53 (*N* = 297, *p* < .001)	.48 (*N* = 295, *p* < .001)
Essay order 3 (cross-validation)	.54 (*N* = 4,162, *p* < .001)	.59 (*N* = 176, *p* < .001)	.55 (*N* = 105, *p* < .001)
Essay order 4 (cross-validation)	.51 (*N* = 3,284, *p* < .001)	.44 (*N* = 116, *p* < .001)	.53 (*N* = 168, *p* < .001)

Table 33

Prediction of Cognitively Based Assessments of, for, and as Learning (CBAL) Pilot Scores by the Sentence Complexity Feature

Criterion	Form K (*N* = 39)	Form M (*N* = 22)
Strand I long essay score	.38 (*p* = .003)	.47 (*p* = .028)
Strand II long essay score	.45 (*p* = .001)	.52 (*p* = .013)
Strand III long essay score	.33 (*p* = .005)	.47 (*p* = .026)
Total Strand I score	.33 (*p* = .015)	.25 (*p* = .26)
Total Strand II score	.47 (*p* = .004)	.44 (*p* = .04)
Total Strand III score	.52 (*p* = .036)	.41 (*p* = .06)
Total score	.40 (*p* = .01)	.40 (*p* = .07)

4.3 Applications and Future Directions

The results reported here are by their nature very tentative, reflecting an exploration of one large dataset (the Attali & Powers [2008] dataset) and its application in a number of small human-scored datasets focused on the CBAL pilot data. Despite these limitations, the models we have built are very suggestive. The analysis suggests the following:

1 That there may be some differences in construct measurement between CBAL pilot strand scores, despite their strong mutual correlations, which if confirmed by future analyses might indicate that humans are able to capture some distinct aspects of each strand

2 That automated scoring can be provided effectively for Strands I and II of the current competency model

3 That it may be worth considering an alternative method of aggregating the scoring, in which fluent mastery of written English, accuracy in sentence-level text production, and the ability to plan and elaborate complex documents are the primary constructs open to automated scoring.

In order to apply and validate such an analysis, and to decide among alternatives, a much larger dataset of scored CBAL pilot tests will need to be developed. But the results are encouraging and suggest that it is quite feasible to plan for automated scoring of CBAL writing (except for Strand III, where human scoring may be more construct-appropriate).

If, however, we take seriously the results of the factor analysis, we may wish to separate the scoring of the two sub-nodes of Strand I and adopt a scoring model in which each of the three factors from that analysis are treated separately. In that light, it is worth considering the scoring rubrics that were applied to CBAL pilots on Forms K and M, and then considering how they might be revised in the light of the results of this study.

The generic rubric used to score essays for Strand I of the competency model is presented as the first table in Appendix B.[29] The results of the factor analyses suggest that the human scoring based upon this rubric may have been more sensitive to grammar and mechanics errors than to students' general maturity in their use of an academic vocabulary and style; at least, the mechanics, usage, and grammar features were more strongly predictive of Strand I human scores than the academic orientation features. It might therefore be very helpful to separate errors in following written conventions from mastery of academic English vocabulary and style, and to score these separately, since the factor analysis indicates that there should be students with good general language skills but significant grammar and mechanics weaknesses, and vice versa.

The factor analysis also suggests that references to sentence structure as in this rubric may be misleading. Some aspects of sentence structure and complexity, in particular those that lead to longer sentences being produced, correlate most tightly with the elaboration factor, and may reflect a general ability to produce structured text. This suggests, in turn, that the Strand II rubric, as shown in Appendix B, also needs revision. Proposed revisions to the rubrics of human scoring, to be used to define the constructs to support automated scoring, are also given in Appendix B.

The analysis reported to date is essentially exploratory, though the exploration does suggest a number of potentially fruitful directions for future research. The potential of timing features to yield additional measurements of text production skill is perhaps the most important of these. We expect in future studies to collect very large samples of student timing data, which will provide more definitive information about the usefulness of the additional features considered in this study.

It is also worth considering that a simpler approach to scoring the CBAL essays may be in order, particularly if larger-scale studies currently underway indicate that there is little statistical separability among the traits considered in the CBAL competency model. It is highly probable, based upon our results to date, that an e-rater–like model could be used to provide automated essay scores trained on a human rubric focused on organization, development, vocabulary, grammar, usage, mechanics, and style. Given this, we might be able to consider a scoring model in which human and machine scoring served different functions. The e-rater score would focus on text production skills, and a cross-checking human score would focus on critical thinking skills. If larger studies confirm a high correlation between the two, the overall essay score could be presented as a composite of the human and machine scores, with a second rater called in only where the two disagreed, to confirm that the difference in scores actually reflected a difference in the quality of text features versus quality of the underlying content.[30] At this point, this possibility must be viewed largely as a speculation into fruitful future directions, and many complications would have to be considered and resolved before such an approach could be adopted. It would, however, provide one way to leverage automated scoring while preserving the commitment that the CBAL approach to writing has made to the importance of rhetorical, content, and critical-thinking based elements to the writing construct and hence to writing instruction.

References

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

Attali, Y., & Powers, D. E. (2008). *A developmental writing scale* (ETS Research Rep. No. RR-08-19). Princeton, NJ: ETS.

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment, 6. Retrieved* from http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1089&context=jtla.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York, NY: Springer.

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

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

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

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

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., …Urzua, A. (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: ETS.

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

Bloom, B. S. (1956). *Taxonomy of educational objectives, Handbook 1: The cognitive domain*. New York, NY: Addison Wesley Publishing Company.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B. and Kukich, K., (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system*. Princeton, NJ: ETS.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine, 25*(3), 27–36.

Burstein, J., & Shermis, M. D. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–122). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213–238.

Crowhurst, M. (1983). Syntactic complexity and writing quality: A review. *Canadian Journal of Education, 8*(1), 1–16.

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 difficulty of reading materials for adults of limited reading ability. *The Library Quarterly, 4*(3), 384–412.

Deane, P., Fowles, M. E., Baldwin, D., & Persky, H. R.,. (2011). *The CBAL summative writing assessment: An 8th grade design* (ETS Research Memorandum No. RM-11-01). Princeton NJ: ETS.

Deane, P., Quinlan, T., Odendahl, N., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive models of writing: Writing proficiency as a complex integrated skill. CBAL literature review-writing* (ETS Research Report No. RR-08-55). Princeton, NJ: ETS.

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.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233.

Flesch, R. (1974). *The Art of Readable Writing*. New York, NY: Harper & Row.

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

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools*. New York, NY: Carnegie Corp.

Hillocks, G. (1987). Synthesis of research on teaching writing. *Educational Leadership, 44*(8), 71.

Hillocks, G. (2002). *The testing trap*. New York, NY: Teachers College Press.

Hunt, K. W. (1970). Syntactic maturity in schoolchildren and adults. *Monographs of the Society for Research in Child Development, 35*(1), iii–67.

Just, M., Carpenter, P. (1971). Comprehension of negation with quantification*. Journal of Verbal Learning and Verbal Behavior, 10*, 244–253.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The intelligent essay assessor. *IEEE Expert Intelligent Systems & Their Applications, 15*, 27.

Langer, J. A. (2001). Beating the odds: Teaching middle and high school students to read and write well. *American Educational Research Journal, 38*(4), 837–880.

Lee, D. (2000). *Modelling variation in spoken and written language: The multi-dimensional approach revisited.* , Lancaster, United Kingdom: Lancaster University.

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* (NCTE Committee on Research Report No. 18). Urbana, IL: National Council of Teachers of English.

Louwerse, M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004, August). *Variation in language and cohesion across written and spoken registers.* Paper presented at the proceedings of the 26th annual conference of the Cognitive Science Society, Chicago, IL.

Matsuhashi, A. (1981). Pausing and planning: The tempo of written discourse production. *Research in the Teaching of English, 15*(2), 113–134.

McNamara, D.S., Ozuru, Y., Graesser, A.C., & Louwerse, M. (2006). Validating Coh-Metrix. In R. Sun, & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 573–578). Austin, TX: Cognitive Science Society.

Meyer, B., Middlemiss, W., Theodorou, E., Brezinski, K., McDougall, J., & Bartlett, B. (2002) Effects of structure strategy instruction delivered to fifth-grade children using the internet with and without the aid of older adult tutors*. Journal of Educational Psychology, 94*, 486-519.Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Monaghan, W., & Bridgeman, B. (2005). E-rater as a quality control on human scores. *R&D Connections, 2*.

O'Reilly, T., Fowles, M. E., Sheehan, K. M., Deane, P., Baldwin, D., & Nadelman, H. (2009). *The CBAL formative assessment of literacy skills*. Unpublished manuscript.

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). The imminence of grading essays by computer. *Phi Delta Kappan, 47*(1), 238–243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*(2), 127–142.

Page, E. B. (1995, August). *Computer grading of essays: A different kind of testing*. Paper presented at the annual meeting of the American Psychological Association, Miami, FL.

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

Pennebaker, J. W., & Francis, M. E. (1999). Linguistic inquiry and word count: LIWC (version 2.0) [Computer software]. Mahwah, NJ: LEA Software and Alternative Media.

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct coverage of the e-rater scoring engine* (ETS Research Rep. No. RR-09-01). Princeton, NJ: ETS.

Rijmen, Frank. 2008. Exploratory statistical results for the spring 2008 CBAL writing periodically administered assessment.   Internal ETS project report.

Schilperoord, J. (2002). On the cognitive status of pauses in discourse production. In T. Olive & C. M. Levy (Eds.), *Contemporary tools and techniques for studying writing* (pp. 61–90). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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 Report No. RR-06-24). Princeton, NJ: ETS.

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

Sheehan, K. M., Kostin, I., & Futagi, Y. (2007b). *Supporting efficient, evidence-centered item development for the GRE paragraph reading item type* (GRE Board Repoprt No. 13-14) Princeton, NJ: ETS.

Sheehan, K. M., Kostin, I. Futagi, Y., & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards* (ETS Research Report RR-10-28). Princeton, NJ: ETS.

Shepard, L. A. (2002). The role of assessment in a learning culture. In C. Desforges & R. Fox (Eds.), *Teaching and learning: The essential readings* (pp. 229–253). Malden, MA: Blackwell.

Stone, P.., Dunphy, D., Smith, M., & Ogilvy, D. (1966) *The general inquirer: A computer approach to content analysis.* Cambridge, MA: MIT Press.

Thorndike, E. L. (1921). Word knowledge in the elementary school. *Teachers College Record, 22*(4), 334–370.

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

Wiliam, D. (2007). Keeping learning on track: Formative assessment and the regulation of learning. In F. K. Lester (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053–1098). Greenwich, CT: Information Age.

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

Notes

[1] Thus there are obvious connections to Bloom's (1956) taxonomy of cognitive skills. This model's focus differs from that taxonomy in that it focuses attention specifically on the writing process and at three levels of abstraction: basic text production and comprehension, the process of producing a structured document, and the critical thinking needed to adapt these processes to specific tasks and rhetorical purposes.

[2] Note that this competency model implies a strong connection between reading and writing. This connection has strong implications for the design of assessments; among other things, it suggests that it may not be reasonable to attempt to assess writing for individuals who are significantly challenged in their reading skills. It may be appropriate in consequence to use screening tasks to identify individuals who are not prepared to tackle an extensive writing task. An example of such a screening task may be found in one of the draft assessments ETS has recently been piloting. The final writing task on this test focuses on producing a persuasive essay. One of the preliminary tasks in the assessment is a simple classification task, in which students are asked to classify 10 statements as making an argument *for* or *against* the issue in focus. Students who are accurate in this task (at least 8 correct out of 10) may perform well or poorly on the essay, but virtually no students who fell below this performance level were able to write a satisfactory persuasive essay.

[3] See Bennett and Gitomer (2009) for related and antecedent thinking about the uses of assessment *for* and *as* learning.

[4] In fact, the assessments built for CBAL are envisaged as forming part of a coordinated set of summative and formative assessments, supported by teacher professional development materials and a large library of formative task sets. The description provided thus far is intended as a description of the summative portion of this design.

[5] The use of task sets that are thematically connected is potentially a source of conflict, since there can be dependencies across tasks that may be useful for instructional purposes but problematic for assessments. We have endeavored to avoid dependencies by designing tasks so that students are provided information designed to bring them to a common level of understanding before they are required to move on to a later task in the set.

[6] In principle, teachers could be used to provide relatively quick human scoring. This may be appropriate in some applications of CBAL, since the training required for teacher scoring might provide effective teacher professional development. However, for many purposes, such as fast turnaround for formative purposes, automated scoring has significant advantages. In the long run it might be best to use automated scoring to deal with some parts of the writing construct for which automated analysis is most effective and to reserve human scoring for those aspects of the writing construct that cannot directly be measured by automated means.

[7] E-rater is by origin highly driven by empirical data. For construct analysis, see Quinlan, Higgins, and Wolff (2009).

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

[9] 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.

[10] We also conducted a factor analysis consisting solely of e-rater microfeatures (features always aggregated with other features to define the features used in the e-rater scoring system). This analysis indicated a relatively small number of factors, with only two emerging consistently and robustly across a variety of conditions. These factors (orthographic accuracy and verb errors) also emerged when the e-rater microfeatures were folded into the global factor analysis.

[11] A set of confirmatory factor analyses was also performed, whose results generally confirmed the second order factor analysis. The most favored factor structures involved no more than three factors.

[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 of course 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 both 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] Note a complication implied by this account. If these features are viewed as we suggest, the construct for Strand I is being able fluently to produce grammatically correct English text using typical written vocabulary and style. This raises questions about prompts for which an oral style is appropriate, and for which it might be inappropriate to focus on features measuring the academic pole of the spoken-to-academic dimension of text variation. It may be necessary to consider this question on a prompt by prompt basis and modify expectations based upon the extent to which a prompt invites or at least allows users to use a particular style; a high level of competency would imply appropriate shifts in register and style, including the capability of using academic style and vocabulary where called for.

[14] For example, some raters were told that some of these essays were written by students two grade levels above or below their actual level and were collected only for a subset of prompts.

[15] In fact, for this data set, the current set of e-rater features produced regression models that predicted human score as much as .10 r-squared higher in performance than regression models that made use of factor scores alone, with the e-rater development and organization features accounting for most of the increased performance. Despite their high correlations with document length, these features appear to capture significant variance that cannot easily be replicated using features not sensitive to discourse representations of the text.

[16] Frank Rijmen, internal CBAL project report, March 2008.

[17] Partly for consistency, all data in the analyses below are given for the datasets for which all features were defined. Operational issues prevented keystroke data being collected for a significant number of students administered Form M, and thus the actual number of responses entered in the analysis for Form M was quite small (22 essays), in contrast to a larger set for Form K (40 essays). The discarded data was collected under quite stressful conditions, requiring students to copy screenshots onto regular word processing software due to network operations issues that negatively impacted the test delivery system.

[18] One of the results for Form K could be less than ideal. The style and usage features load on both the Strand I and the Strand II models, which could reflect an inappropriate aggregation of microfeatures. In particular, the style feature was already shown in Section 2 to be questionable, involving as it does a combination of features loading in opposite directions on at least two underlying factors. We therefore constructed regression models in which a

normalized version of the repetition of words feature was added to the stepwise regression, along with several features associated with the sentence length factor. In Form K, where the style feature played a role, it was replaced in both Strands I and II by the repetition-of-words microfeature, its most important component feature, without any significant overall improvement to the model. This suggests that the other features added little prediction or even some noise, at least for this dataset, but that the repetition-of-words feature is itself ambiguous, reflecting aspects both of Strand I and of Strand II. For Test M, both the style and repetition-of-words features failed to make it into the Strand I model, but a feature characteristic of short sentences (a high proportion of verbs in the sentence) was added, bringing the r-square up to .77, an increase of .07 over the model that made use of the grammar and mechanics features alone. These results do not demonstrate that the style feature should not be aggregated the way it is, but they raise the possibility rather strongly, at least for predicting CBAL strand scores. However, no similar result was obtained for the usage feature, although most of its component features are extremely sparse.

[19] Pauses between paragraphs were also recorded but not used in the analysis, as the data was very sparse, since a large proportion of students produced zero or only one break between paragraphs in their responses.

[20] Russell Almond (internal CBAL project report) indicated that most of these features can be modeled statistically as a mixture of log-normals with two or at most three components (e.g., a short-pause component and a longer-pause component). It is possible that an analysis could be developed based upon such components, which might reflect differences (e.g., between motor and planning processes). This possibility will be examined in greater depth in future work.

[21] Also, longer breaks between sentences helped to predict higher scores on a regression model for Strand III on Form M but not on Form K.

[22] Of course, reporting at a student level presents even more challenges, which will not be addressed here, since the information must be translated into a comprehensible form and wedded to activities that will drive student engagement and learning if it is to significantly affect student performance.

[23] This assignment is not clear. Sentence complexity is related not only to syntactic variety, and thus to using a more academic style, but also to the general ability fluently to produce structured documents, which corresponds to the plan/structure document node. The factor analysis favors an interpretation in which we view sentence complexity as measuring pretty much the same complex of abilities as document length and the organization and development features, which could be viewed as a pure fluency factor but which, we argue, reflects the ability to produce structured texts.

[24] This assignment is also somewhat problematic, in that grammar and usage errors could reflect dialect differences or production errors induced by stress, quite separate from the ability to proofread and correct such errors when they are produced. Timing features might help to disambiguate such features, but for now we assign them with the caveats noted.

[25] It is also encouraging that this feature has only a .29 correlation with document length in the Attali and Powers (2008) dataset, far lower than most e-rater features, though slightly higher than the .20 correlations for word length and median word frequency.

[26] Excluding outliers where students produced less than 10 words total.

[27] The equations use a variety of features available from the underlying SourceFinder-derived feature set, many of which are closely correlated. By using multiple regression equations, it was possible to approximate an analysis in which these features jointly determine the predicted variable. The equations were calculated factoring out important competing variables such as written vocabulary and document length. Since all the features used were calculated similarly (normalizing by document length and taking the log), they are roughly on the same scale.

[28] However, if document length is factored out, only long essay strand I scores have a significant (p <.05) correlation with this feature in the CBAL pilot datasets. Note that the written style feature has a correlation of .40 with document length, which is much lower than that observed with many e-rater features.

[29] In the lower half, the diagnostic letters a, b, c, etc., were used to indicate particular categories of error thought likely to be of interest for reporting purposes. Fall pilot data suggests that

more work needs to be done to make these error categories reliable enough to report (Rijmen, 2008).

[30] A similar idea is presented by Monaghan & Bridgeman (2005), which presented use of e-rater as a means of quality-checking human scores. What we are proposing is somewhat different, as we anticipate that automated features could be trained to a level where they would be fairly reliable indicators of those aspects of performance that they directly measured, so that the human score could be used as a subscore specifically intended to measure such aspects of the construct as rhetorical effectiveness and quality of reasoning.

List of Appendices

Appendix A

Feature Sources

The features considered here were used primarily for the convenience: They have already been developed and computationally implemented and thus can be assessed and combined without a long feature development process. However, the actual performance of each feature when embedded in an essay scoring context may need additional research, and it may in future be necessary to develop features designed to measure aspects of the competency model not fully covered in the current, convenience, set of features. See Quinlan et al. (2009) for more detail on performance issues with specific e-rater features.

Table A1

Feature Sources

Feature name	Source (though generally modified)
Abstract nouns	Biber et al. (2004)
Academic verbs	Biber et al. (1999)
Cognitive process perception nouns	Biber et al. (2004)
Adversative conjunctions	Louwerse et al. (2004)
Clarifying conjunctions	Louwerse et al. (2004)
Contractions	Biber (1988)
Indefinite pronouns	Biber (1988)
Possibility modals	Biber (1988)
Prediction modals	Biber (1988)
Narrative communication verbs	Biber et al. (2004)
Narrative mental state verbs	Biber et al. (2004)
Sentence negation	Biber (1988)
Nominalizations	Biber (1988), Lee (2000)
Prepositions	Biber (1988)
First person pronouns	Biber (1988)
Second person pronouns	Biber (1988)
Third person pronouns	Biber et al. (1999)
Causal subordinator	Biber (1988)
Concessive subordinator	Biber (1988)
Multifunctional subordinator	Biber (1988)
Fiction verbs	Biber et al. (1999)
Conversation verbs	Biber et al. (1999)
Demonstratives	Biber (1988)
Verbs of causation	Biber et al. (2004)
Meyer causative list	Meyer et al. (2002)
Causal particles	McNamara et al. (2006)
Academic downtoners	Biber (1988)

Feature name	Source (though generally modified)
Relational adjectives	Biber et al. (2004)
Adverbials of place	Biber (1988)
Adverbials of time	Biber (1988)
Belief words	e-rater
Causal conjunctions	Louwerse et al. (2004)
Exclamation marks	Flesch (1974)
Necessity modals	Biber (1988)
Negative prefixes	Just et al. (1971)
Research words	Sheehan et al. (2007b)
LIWC causal words	Pennebaker & Francis (1999)
LIWC inclusive words	Pennebaker & Francis (1999)
Stone gender words	Stone et al. (1966)
Flesch group words	Flesch (1974)
LIWC human words	Pennebaker & Francis (1999)
Coxhead academic word list	Coxhead (2000)
Emotion words	Stone et al. (1966)
Activity words	Biber et al. (2004)

Appendix B

Additional Tables (e-rater Microfeatures)

These tables are summarized in the discussion in the body of the text. It is worthwhile to present them in detail here, since the analysis depends critically upon the overall picture that they provide.

Table B1

Correlations Between e-rater Features and Factor Scores, Essay Order 1

	Grammar	Usage	Mechanics	Style	Organization	Development	Median word frequency	Average word length
Academic orientation	0.27	0.20	0.15	**0.47**	0.29	0.05	**-0.79**	**0.88**
Noun-centered text	0.14	-0.13	0.02	**0.51**	0.18	0.12	**-0.59**	**0.55**
Sentence complexity	**0.32**	0.12	0.14	**0.37**	0.12	0.35	-0.10	0.11
Elaboration	**0.44**	0.27	0.27	**0.39**	**0.41**	0.26	0.19	-0.14
Spoken style	-0.04	0.01	0.08	-0.29	-0.09	0.01	**0.61**	**-0.55**
Overt expression of persuasion	0.11	-0.02	0.01	0.08	0.10	0.07	0.05	0.00
Orthographic accuracy	**-0.31**	**-0.43**	**-0.78**	-0.15	-0.20	-0.04	0.14	-0.10
Narrative style	0.19	0.24	0.07	0.13	0.13	0.13	-0.03	-0.09
Verb errors	**-0.34**	-0.23	-0.20	-0.08	0.01	0.02	0.25	-0.28
Comma errors	0.03	-0.10	-0.11	0.22	0.10	0.01	-0.08	0.18

Note. For convenience, the values that have strong weights on particular factors are in boldface to make them easier to identify.

Table B2

Correlations Between e-rater Features and Factor Scores, Essay Order 2

	Grammar	Usage	Mechanics	Style	Organization	Development	Median word frequency	Average word length
Academic orientation	**0.30**	0.24	0.17	**0.50**	0.32	0.06	-0.79	0.87
Noun-centered text	0.19	-0.06	0.04	**0.54**	0.26	0.11	-0.59	0.58
Sentence complexity	**0.31**	0.11	0.13	**0.36**	0.12	0.33	-0.10	0.12
Elaboration	**0.37**	0.21	0.22	**0.33**	**0.37**	0.22	0.20	-0.17
Spoken style	-0.08	-0.03	0.03	-0.25	-0.12	0.01	0.60	-0.51
Overt expression of persuasion	0.10	0.00	0.05	0.09	0.13	0.05	-0.01	0.06
Orthographic accuracy	**-0.31**	**-0.37**	**-0.77**	-0.16	-0.20	-0.01	0.18	-0.17
Narrative style	0.16	0.25	0.02	0.05	0.08	0.09	-0.04	-0.11
Verb errors	**-0.40**	-0.28	-0.24	-0.03	-0.02	0.01	0.17	-0.12
Comma errors	0.00	-0.17	-0.11	0.18	0.07	0.01	-0.02	0.16

Note. For convenience, the values that have strong weights on particular factors are in boldface to make them easier to identify.

Table B3

Correlations Between e-rater Features and Factor Scores, Essay Order 3

	Grammar	Usage	Mechanics	Style	Organization	Development	Median word frequency	Average word length
Academic orientation	**0.31**	0.21	0.16	**0.53**	**0.31**	0.07	**-0.77**	**0.86**
Noun-centered text	0.17	-0.11	-0.02	**0.55**	0.22	0.11	**-0.56**	**0.56**
Sentence complexity	**0.34**	0.08	0.12	**0.37**	0.15	**0.32**	-0.04	0.12
Elaboration	**0.42**	0.24	**0.30**	**0.31**	**0.39**	0.21	0.19	-0.14
Spoken style	-0.03	-0.03	0.04	-0.22	-0.08	0.04	**0.58**	**-0.50**
Overt expression of persuasion	0.12	-0.03	0.04	0.03	0.13	0.02	0.07	-0.04
Narrative style	0.16	0.25	0.05	0.10	0.08	0.11	-0.10	-0.03
Orthographic accuracy	-0.28	**-0.32**	**-0.73**	-0.17	-0.20	0.00	0.12	-0.15
Verb errors	**-0.37**	**-0.35**	**-0.32**	0.04	0.00	0.05	0.15	-0.14
Comma errors	-0.07	-0.21	-0.28	0.14	0.03	-0.01	0.05	0.09

Note. For convenience, the values that have strong weights on particular factors are in boldface to make them easier to identify.

Table B4

Correlations Between e-rater Features and Factor Scores, Essay Order 4

	Grammar	Usage	Mechanics	Style	Organization	Development	Median word frequency	Average word length
Academic orientation	**0.30**	0.24	0.17	**0.50**	0.32	0.06	**-0.79**	**0.87**
Noun-centered text	0.19	-0.06	0.04	**0.54**	0.26	0.11	**-0.59**	**0.58**
Sentence complexity	**0.31**	0.11	0.13	**0.36**	0.12	0.33	-0.10	0.12
Elaboration	**0.37**	0.21	0.22	**0.33**	0.37	0.22	0.20	-0.17
Spoken style	-0.08	-0.03	0.03	-0.25	-0.12	0.01	**0.60**	**-0.51**
Overt expression of persuasion	0.10	0.00	0.05	0.09	0.13	0.05	-0.01	0.06
Orthographic accuracy	**-0.31**	**-0.37**	**-0.77**	-0.16	-0.20	-0.01	0.18	-0.17
Narrative style	0.16	0.25	0.02	0.05	0.08	0.09	-0.04	-0.11
Verb errors	**-0.40**	-0.28	-0.24	-0.03	-0.02	0.01	0.17	-0.12
Comma errors	0.00	-0.17	-0.11	0.18	0.07	0.01	-0.02	0.16

Note. For convenience, the values that have strong weights on particular factors are in boldface to make them easier to identify.

Table B5

Correlations Between e-rater Style Features and the 10 Factors (Across All Essay Orders)

	Academic orientation	Noun-centered text	Sentence complexity	Spoken style	Persuasive style	Elaboration	Narrative style	Ortho-graphic accuracy	Verb errors	Comma errors
GUMS 401 (repetition of words)	-.40 (p < .001)	-.46 (p < 0.001)	-.26 (p < .001)	.31 (p < .03)	-.36 (p < .001)	-.03 (p < .07)	-.11 (p < .001)	.13 (p < .001)	**.06** **(p < .001)**	**-.21** **(p < .001)**
GUMS 402 (inappro-priate word or phrase)	-.01 (p < .53)	.01 (p < .35)	.00 (p < .80)	-.01 (p < .46)	.03 (p < .05)	.00 (p < .99)	.03 (p < .02)	.03 (p < .06)	**.00** **(p < .98)**	**-.01** **(p < .58)**
GUMS 403 (excessive use of and)	-.07 (p < .001)	-.04 (p < .01)	-.02 (p < .10)	.03 (p < .02)	.04 (p < .001)	.01 (p < .31)	.03 (p < .06)	.06 (p < .02)	**.05** **(p < .001)**	**-.03** **(p < .03)**
GUMS 404 (too many short sentences)	-.15 (p < .001)	-.15 (p < .001)	-.38 (p < .001)	.06 (p < .001)	.06 (p < .001)	-.11 (p < .001)	.04 (p < .002)	.01 (p < .63)	**.08** **(p < .001)**	**-.03** **(p < .04)**
GUMS 405 (too many long sentences)	-.03 (p < .02)	.01 (p < .51)	.33 (p < 0.001)	.05 (p < .002)	.09 (p < .001)	.07 (p < .001)	.05 (p < .001)	.07 (p < .001)	**.04** **(p < .001)**	**.01** **(p < .45)**
GUMS 406 (overuse of agentive passives)	0.18 (p < .001)	.15 (p < .001)	.06 (p < .001)	-.15 (p < .001)	.09 (p < .001)	-.08 (p < .001)	.07 (p < .001)	-.09 (p < .001)	**-.03** **(p < .03)**	**.01** **(p < .63)**

Note. GUMS =grammar, usage, mechanics, and style.

Table B6

Correlation Between Human Score/Grade and Normalized e-rater Micro-Features for e-rater Style Features (All Essay Orders)

	Persuasive essays (human score) *N* = 589	Descriptive essays (human score) *N* = 568	Grade level *N* = 17,536
GUMS 401 (repetition of words)	-.50 *p* < .001	-.48 *p* < .001	-.32 *p* < .001
GUMS 402 (inappropriate word or phrase)	N/A	.07 *p* < .10	.02 *p* < .003
GUMS 403 (excessive use of and)	-.02 *p* < .59	.07 *P* < .11	-.03 *P* < .001
GUMS 404 (too many short sentences)	-.02 *p* < .59	.07 *P* < .11	-.11 *P* < .001
GUMS 405 (too many long sentences)	-.01 *p* < .83	.03 *p* < .43	.02 *P* < .01
GUMS 406 (overuse of agentive passives)	.16 *p* < .001	.20 *p* < .001	.17 *p* < .001

Note. All features calculated as log rate per thousand words. GUMS =grammar, usage, mechanics, and style.

Table B7

Correlation Between Human Score/Grade and Normalized e-rater Microfeatures for e-rater Style Features (by Essay Order)

	Persuasive				Descriptive				Grade level			
	Essay order 1	Essay order 2	Essay order 3	Essay order 4 *N* = 116	Essay order 1	Essay order 2	Essay order 3	Essay order 4 *N* = 168	Essay order 1	Essay order 2	Essay order 3	Essay order 4 *N* = 3,284
GUMS 401 (repetition of words)	-.50 *p* < .001	-.55 *p* < .001	-.51 *p* < .001	-.39 *p* < .001	-.51 *p* < .001	-.47 *p* < .002	-.43 *p* < .001	-.53 *p* < .001	-.35 *p* < .001	-.33 *p* < .001	-.31 *p* < .001	-.29 *p* < .001
GUMS 402 (inappropriate word or phrase)	n/a	n/a	n/a	n/a	n/a	n/a	.17 *p* < .09	.05 *P* < .57	.04 *p* < .01	.01 *p* < .75	.03 *p* < .04	.02 *p* < .26
GUMS 403 (excessive use of and)	-.10 *p* < .32	-.13 *p* < .08	.00 *p* < .97	n/a	.03 *p* < .75	.02 *p* < .82	n/a	.05 *P* < .57		-.03 *p* < .09	-.04 *P* < .03	-.01 *P* < .56
GUMS 404 (too many short sentences)	-.06 *p* < .56	-.15 *p* < .05	.13 *p* < .10	.09 *p* < .36	.15 *p* < .05	.06 *p* < .49	.04 *p* < .72	.04 *P* < .66	-.11 *p* < .001	-.11 *p* < .001	-.11 *p* < .001	-.09 *p* < .001
GUMS 405 (too many long sentences)	-.03 *p* < .75	.05 *p* < .55	.08 *p* < .29	.23 *p* < .02	.10 *p* < .22	-.13 *p* < .09	-.04 *p* < .72	.01 *P* < .86	.02 *P* < .13	.03 *p* < .06	.08 *p* < .29	.05 *p* < .01
GUMS 406 (overuse of agentive passives)	.16 *p* < .10	.19 *p* < .01	.13 *p* < .09	.18 *p* < .06	.13 *p* < .11	.07 *p* < .40	.09 *p* < .36	.37 *P* < .001	.16 *P* < .001	.17 *p* < .001	.17 *p* < .001	.17 *P* < .001

Note. All features calculated as log rate per thousand words. GUMS = grammar, usage, mechanics, and style.

Table B8

Correlation Between the 10 Factors and Normalized e-rater Microfeatures for e-rater Mechanics Features (All Essay Orders)

	Academic orientation	*N*oun-centered text	Sentence complexity	Spoken style	Persuasive style	Elaboration	*N*arrative style	Orthographic accuracy	Verb errors	Comma errors
GUMS 301 (spelling)	-.05 $p < .001$	.06 $p < .001$	-.13 $p < .001$	-.18 $p < .001$	-.26 $p < .001$	.01 $p < .33$	-.05 $p < .001$	.67 $p < .001$	.21 $p < .001$	.11 $p < .001$
GUMS 302 (didn’t capitalize proper noun)	-.21 $p < .001$	-.15 $p < .001$	.01 $p < .44$	.19 $p < .001$	.07 $p < .001$	-.06 $p < .001$	.11 $p < .001$	.57 $p < .001$	.03 $p < .06$	-.18 $p < .001$
GUMS 304 (missing question mark)	-.04 $p < .007$	-.02 $p < .19$	-.05 $p < .001$	.03 $p < .03$	-.02 $p < .09$	.01 $P < .39$	-.01 $P < .41$	.06 $p < .001$	.06 $p < .001$	.01 $p < .48$
GUMS 306 (absent apostrophes in contractions)	-.14 $p < .001$	-.07 $p < .001$	-.08 $p < .001$	.17 $p < .001$	.05 $p < .001$	.08 $p < .001$	-.11 $p < .001$	.66 $p < .001$	.05 $p < .001$	.08 $p < .001$
GUMS 307 (comma errors)	-.02 $p < .19$	-.02 $p < .14$	.03 $p < .01$	.03 $p < .05$	.02 $p < .15$	.001 $p < .73$	.01 $p < .46$	.05 $p < .001$	.06 $p < .001$	.736 $p < .001$
GUMS 308 (hyphen errors)	.05 $p < .001$	.10 $p < .001$	.03 $p < .03$	-.06 $p < .001$	.03 $p < .04$	-.02 $p < .24$	.06 $p < .001$	.01 $p < .66$	-.02 $P < .26$	.02 $P < .26$
GUMS 309 (one word should be two)	-.01 $p < .5$	-.03 $p < .06$	.03 $p < .03$	.03 $p < .04$	.01 $p < .66$	-.01 $p < .40$	.01 $p < .73$	.16 $p < .001$	.07 $p < .001$	.02 $p < .21$
GUMS 310 (two words should be one)	-.07 $p < .001$	-.01 $p < .45$	.04 $p < .001$	.02 $p < .11$	.02 $p < .09$	.03 $p < .02$	-.05 $p < .001$	.09 $p < .001$	.09 $p < .001$	.04 $P < .004$
GUMS 311 (two adjacent identical words)	-.06 $p < .001$	-.01 $p < .36$	-.01 $p < .67$	.02 $p < .10$	.03 $p < .08$	.04 $p < .004$	.01 $p < .72$	.08 $p < .001$	.06 $p < .001$	-.01 $p < .40$

Note. GUMS = grammar, usage, mechanics, and style.

Table B9

Correlation Between Human Score/Grade and Normalized e-rater Microfeatures for e-rater Mechanics Features (All Essay Orders)

	Persuasive essays (human score) N =589	Descriptive essays (human score) N =568	Grade level N =17,536
GUMS 301 (spelling)	-.08 $p < .06$	-.21 $p < .001$	-.18 $p < .001$
GUMS 302 (didn’t capitalize proper noun)	-.12 $p < .005$	-.08 $p < .06$	-.04 $p < .001$
GUMS 304 (missing question marks)	-.04 $p < .38$	-.08 $p < .05$	-.03 $p < .001$
GUMS 306 (absent apostrophes in contractions)	-.08 $p < .07$	-.06 $p < .17$	-.06 $p < .001$
GUMS 307 (comma errors)	-.03 $p <49$	-.01 $p < .85$	.00 $p < .77$
GUMS 308 (hyphen errors)	.15 $p < .001$	.07 $p < .12$	.06 $p < .001$
GUMS 309 (one word should be two)	-.04 $p < .33$	-.12 $p < .004$	-.03 $p < .001$
GUMS 310 (two words should be one)	.03 $p < .48$	-.01 $p < .88$	.03 $p < .001$
GUMS 311 (two adjacent identical words)	-.04 P\ $p < .39$	-.01 $p < .88$	-.03 $p < .001$

Note. All features calculated as log rate per thousand words. GUMS = grammar, usage, mechanics, and style.

Table B10

Correlation Between Human Score/Grade and Normalized e-rater Microfeatures for e-rater Mechanics Features (by Essay Order)

	Persuasive				Descriptive				Grade level			
	Essay	Essay	Essay	Essay	Essay	Essay	Essay	Essay	Essay	Essay	Essay	Essay
	order 1	order 2	order 3	order 4	order 1	order 2	order 3	order 4	order 1	order 2	order 3	order 4
	N = 162	*N* = 187	*N* = 176	*N* = 116	*N* = 110	*N* = 133	*N* = 105	*N* = 168	*N* = 5,150	*N* =4,940	*N* =4,162	*N* =3,284
GUMS 301 (Spelling, normalized and logged)	.00 *p* < .98	-.05 *p* < .51	-.15 *p* < .06	-.17 <.07	-.23 *p* < .005	-.21 *p* < .02	-.15 *p* < .12	-.21 *p* < .007	-.17 *p* < .001	-.16 *p* < .001	-.18 *P* < .001	-.22 *P* < .001
GUMS 302 (Didn't capitalize proper noun, normalized and logged)	-.10 *p* < .29	-.15 *p* < .05	-.03 *p* < .75	-.22 *p* < .02	.02 *p* < .85	-.21 *p* < .02	-.03 *p* < .82	-.13 *p* < .09	-.06 *p* < .001	-.06 *p* < .001	-.02 *p* < .18	-.03 *p* < .14
GUMS 304 (Missing question mark, normalized and logged)	-.14 *p* < .16	-.04 *p* < .57	.07 *p* < .35	.01 *p* < .93	.04 *p* < .61	-.17 *p* < .05	-.14 *p* < .17	-.14 *p* < .07	-.03 *p* < .04	-.02 *p* < .10	-.02 *p* < .32	-.05 *p* < .01
GUMS 306 (Absent apostrophes in contractions, normalized and logged)	-.04 *p* < .67	-.15 *p* < .04	-.04 *p* < .62	-.08 *p* < .29	.06 *p* < .45	-.16 *p* < .07	-.08 *p* < .41	-.05 *p* < .62	-.05 *p* < .001	-.06 *p* < .001	-.06 *p* < .001	-.06 *p* < .001
GUMS 307 (comma errors, normalized and logged)	-.06 *p* < .57	.04 *p* < .62	-.07 *p* < .33	-.05 *P* < .58	.03 *p* < .71	.09 *p* < .33	.13 *p* < .20	-.17 *p* < .03	-.01 *p* < .41	.01 *p* < .64	.00 *p* < .83	-.004 *P* < .83
GUMS 308 (hyphen errors, normalized and logged)	.12 *P* < .23	.20 p < .007	.10 *p* < .18	.14 *p* < .15	.12 *p* < .12	-.02 *p* < .80	-.01 *P* < .96	.12 *p* < .12	.05 *p* < .001	.07 *p* < .001	.06 *p* < .001	.09 *p* < .001
GUMS 309 (one word should be two, normalized and logged)	.01 *P* < .89	-.01 *p* < .94	-.05 *P* < .51	-.19 *p* < ..04	.00 *p* < .97	-.06 *p* < .48	-.12 *p* < .22	-.25 *p* < .001	-.01 *p* < .74	-.04 *p* < .02	-.03 *p* < .07	-.05 *p* < .003
GUMS 310 (two words should be one, normalized and logged)	-.08 *p* < .43	.03 *p* < .67	.11 *p* < .13	.01 *p* < .9	-.13 *p* < .10	.02 *p* < .82	-.11 *p* < .27	-.09 *P* < .27	.01 *p* < .64	.02 *p* < .10	.04 *p* < .006	.04 *P* < .03
GUMS 311 (two adjacent identical words, normalized and logged)	-.14 *p* < .16	-.08 *P* < .31	.02 *P* < .78	.11 *P* < .26	-.08 *p* < .32	.18 *p* < .04	.04 *P* < .67	-.07 *P* < .35	-.04 *p* < .002	-.03 *p* < .06	-.02 *P* < .29	-.03 *P* < .009

Note. GUMS = grammar, usage, mechanics, and style.

Table B11

Correlation Between the 10 Factors and Normalized e-rater Microfeatures for e-rater Usage Features Over All Essay Orders

	Academic orientation	Noun-centered text	Sentence complexity	Spoken Style	Persuasive style	Elaboration	Narrative style	Orthographic accuracy	Verb errors	Comma errors
GUMS 201 (wrong article)	0.04 $p < 0.01$	.06 $p < 0.001$	.04 $p < .003$	-.03 $p < .06$	.03 $p < .03$	.01 $p < .55$	-.02 $p < .09$	.06 $p < .001$	.05 $p < .001$	**.06** $p < .001$
GUMS 202 (wrong, missing or confused article)	0.09 $p < 0.001$	.47 $p < 0.001$	.03 $p < .03$	-.18 $p < .003$	.04 $p < .007$	.08 $p < .001$	-.04 $p < .009$	.06 $p < .001$	.15 $p < .001$	**.07** $p < .001$
GUMS 203 (confusion of homonyms)	-0.14 $p < 0.001$	.11 $p < 0.001$	.13 $p < .001$	.04 $p < .005$	.16 $p < .001$	.09 $p < .001$	-.15 $p < .001$	.48 $p < .001$	.30 $p < .001$	**.23** $p < .001$
GUMS 205 (faulty comparison)	0.02 $p < 0.23$	.03 $p < 0.03$	.00 $p < .82$	-.01 $p < .61$	.00 $p < .74$	-.01 $p < .29$	-.03 $p < .07$	.03 $p < .06$	.00 $p < .90$	**.03** $p < .02$
GUMS 207 (nonstandard verb or word form)	-0.03 P<0.02	-.03 $p < 0.5$	.02 $p < .19$	.03 $p < .04$	.01 $p < .39$	.04 $p < .06$	.00 $p < .91$	.07 $p < .001$	.03 $p < .06$	**.00** $p < .98$

Note. GUMS = grammar, usage, mechanics, and style.

Table B12

Correlation Between Human Score/Grade and Normalized e-rater Microfeatures for e-rater Mechanics Features (All Essay Orders)

	Persuasive essays (human score) *N* =589	Descriptive essays (human score) *N* =568	Grade level *N* =17,536
GUMS 201 (wrong article)	.04 $p < .28$	.06 $P < .16$	.04 $p < .001$
GUMS 202 (wrong, missing or confused article)	.11 $P < .005$	.07 $p < .09$	.12 $p < .001$
GUMS 203 (confusion of homonyms)	-.12 $P < .005$	-.13 $p < .002$	-.04 $p < 001$
GUMS 205 (faulty comparison)	.00 $p < .97$	.00 $p < .92$	-.02 $p < .04$
GUMS 207 (nonstandard verb or word form)	.00 $p < .94$	-.01 $p < .76$	-.01 $p < .11$

Note. All features calculated as log rate per thousand words. GUMS = grammar, usage, mechanics, and style.

Table B13

Correlation Between Human Score/Grade and Normalized e-rater Microfeatures for e-rater Usage Features (Broken out by Essay Order)

	Persuasive				Descriptive				Grade level			
	Essay order 1 *N* =162	Essay order 2 *N* =187	Essay order 3 *N* =176	Essay order 4 *N* =116	Essay order 1 *N* =110	Essay order 2 *N* =133	Essay order 3 *N* =105	Essay order 4 *N* =168	Essay order 1 *N* =5,150	Essay order 2 *N* =4,940	Essay order 3 *N* =4,162	Essay order 4 *N* =3,284
GUMS 201 (wrong article)	-.16 *P* < .10	.07 *P* < .35	.15 *p* < .05	.14 *p* < .14	.12 *p* < .13	-.03 *p* < .70	.21 *p* < .04	.002 *p* < .98	.05 *p* < .001	.05 *P* < .001	.03 *p* < .12	.02 *p* < .31
GUMS 202 (wrong, missing or confused article)	.11 *p* < .24	.04 *P* < .64	.21 *P* < .007	.15 *p* < .10	.03 *P* < .73	.07 *P* < .43	.04 *P* < .72	.15 *p* < .05	.13 *p* < .001	.12 *p* < .001	.12 *p* < .001	.11 *p* < .001
GUMS 203 (confusion of homonyms)	-.21 *p* < .003	-.06 *p* < .39	-.06 *p* < .41	-.24 *p* < .01	-.003 *p* < .97	-.13 *p* < .15	-.09 *p* < .39	-.27 *p* < .001	-.03 *p* < .06	-.04 *p* < .02	-.03 *p* < .04	-.06 *p* < 001
GUMS 204	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	.02 *P* < .23	-.01 *p* < .71
GUMS 205 (faulty comparison)	n/a	-.12 *p* < .17	n/a	n/a	.01 *p* < .87	n/a	n/a	n/a	-.004 *p* < .77	-.02 *p* < .10	-.02 *P* < .31	-.03 *p* < .15
GUMS 207 (nonstandard verb or word Form)	-.04 *p* < .66	-.04 *P* < .59	.02 *p* < .79	.08 *p* < .37	.05 *p* < .58	-.12 *p* < .17	.04 *p* < .66	-.02 *p* < .78	-.003 *p* < .84	-.03 *p* < .07	.009 *p* < .52	-.03 *p* < .08

Note. GUMS = grammar, usage, mechanics, and style.

Table B14

Correlation Between the 10 Factors and Normalized e-rater Microfeatures for e-rater Grammar Features (All Essay Orders)

	Academic orientation	Noun-centered text	Sentence complexity	Spoken style	Persuasive style	Elaboration	Narrative style	Orthographic accuracy	Verb errors	Comma errors
GUMS 101 (sentence fragments)	-.06 *p* < .001	.07 *P* < .001	-.20 *p* < .001	-.04 *p* < .001	-.02 *p* < .13	-.01 *p* < .53	.02 *p* < .12	.15 *p* < .001	.05 *p* < .001	.05 *p* < .001
GUMS 103 (garbled sentences)	-.06 *p* < .001	-.01 *p* < .36	.05 *p* < .002	.05 *p* < .001	.04 *p* < .005	.03 *p* < .02	-.02 *p* < .11	.19 *p* < .001	.12 *p* < .001	.05 *p* < .001
GUMS 104 (subject-verb agreement)	-.07 *p* < .001	-.01 *p* < .71	-.12 *p* < .001	.06 *p* < .001	-.06 *p* < .001	.04 *p* < .02	-.08 *p* < .001	.14 *p* < .001	.59 *p* < .001	.11 *p* < .001
GUMS 105 (ill-formed verb)	-.04 *p* < .095	-.02 *p* < .15	.07 *p* < .001	.02 *P* < .23	.06 *p* < .001	.02 *p* < .14	.05 *p* < .001	.02 *p* < .12	.60 *p* < .001	.08 *p* < .001
GUMS 106 (pronoun error)	-.03 *p* < .04	-.01 *p* < 0.53	.00 *P* < .81	.02 *p* < .13	.00 *p* < .73	-.01 *p* < .68	.01 *p* < .72	.04 *p* < .006	.03 *p* < .07	.01 *p* < .33
GUMS 107 (possessive error)	.12 *p* < .001	.14 *p* < .001	.07 *p* < .001	-.11 *p* < .001	.00 *p* < .80	-.03 *p* < .04	.01 *p* <41	.08 *p* < .001	.01 *p* < .71	.07 *p* < .001
GUMS 108 (wrong or missing word)	-.02 *p* < .22	.03 *p* < .04	.04 *p* < .008	.02 *p* < .10	.05 *p* < .001	.00 *p* < .98	-.02 *p* < .25	.02 *p* < .08	.04 *p* < .004	-.07 *p* < .002
GUMS 109 (proofread this)	-.09 *p* < .001	-.03 *p* < 0.04	.05 *P* < .002	-.01 *p* < .32	-.05 *p* < .002	.05 *p* < 0.001	.04 *p* < .005	.11 *p* < .001	.48 *p* < .001	-0.24 *p* < .001

Note. GUMS = grammar, usage, mechanics, and style.

Table B15

Correlation Between Human Score/Grade and Normalized e-rater Microfeatures for e-rater Grammar Features

	Persuasive				Descriptive				Grade level			
	Essay order 1 *N* =162	Essay order 2 *N* =187	Essay order 3 *N* =176	Essay order 4 *N* =116	Essay order 1 *N* =110	Essay order 2 *N* =133	Essay order 3 *N* =105	Essay order 4 *N* =168	Essay order 1 *N* =5,150	Essay order 2 *N* =4,940	Essay order 3 *N* =4,162	Essay order 4 *N* =3,284
GUMS 101 (sentence fragments)	-.19 *p* < .05	-.10 *p* < .19	.09 *p* < .23	-.06 *p* < .56	-.04 *p* < .62	.06 *p* < .52	-.05 *p* < .61	-.05 *p* < .50	-.04 *p* < .01	-.06 *p* < .001	-.06 *P* < .001	-.07 *p* < .001
GUMS 103 (garbled sentences)	-.06 *p* < .55	-.08 *p* < .31	-.05 *p* < .50	-.008 *p* < .94	.03 *p* < .69	-.06 *p* < .48	-.18 *p* < .07	n/a	.02 *p* < .29	-.001 *p* < .94	-.01 *P* < .53	.02 *p* < .37
GUMS 104 (subject-verb agreement)	-.15 *p* < .16	.00 *p* < .99	-.08 *p* < .32	.05 *p* < .61	-.15 *p* < .07	-.12 *p* < .18	-.22 *p* < .03	-.08 *p* < .33	.01 *p* < .56	.006 *p* < .68	-.01 *p* < .50	.03 *p* < .12
GUMS 105 (ill-formed verb)	.02 *p* < .85	.01 *p* < .88	-.02 *p* < .78	.08 *p* < .42	-.14 *p* < .14	.12 *p* < .17	-.002 *p* < .98	-.02 *p* < .78	.02 *p* < .19	.01 *p* < .45	.01 *p* < .54	.004 *p* < .81
GUMS 106 (pronoun error)	.04 *p* < .69	n/a	-.10 *p* < .21	.008 *p* < .94	n/a	n/a	n/a	n/a	-.01 *p* < .52	-.01 *p* < .40	-.02 *p* < .18	-.05 *p* < .006
GUMS 107 (possessive error)	-.05 *p* < .64	.11 *p* < .15	-.11 *p* < .14	-.06 *p* < .54	-.01 *p* < .95	.07 *p* < .46	.06 *p* < .56	-.02 *p* < .76	.09 *p* < .001	.08 *p* < .001	.07 *P* < .001	.09 *p* < .001
GUMS 108 (wrong or missing word)	.01 *p* < .90	-.002 *p* < .98	.00 *p* < .99	-.02 *P* < .87	-.10 *P* < .21	.04 *p* < .62	-.08 *p* < .44	.008 *p* < .93	-.002 *p* < .89	.01 *p* < .60	.01 *p* < .38	-.004 *p* < .81
GUMS 109 (proofread this)	.05 *p* < .60	-.11 *p* < .14	-.04 *p* < .62	-.21 *p* < .03	-.06 *p* < .48	.09 *p* < .30	-.10 *p* < .29	-.16 *p* < .003	.00 *p* < .99	-.03 *p* < .08	-.02 *p* < .11	-.006 *p* < .72

Note. GUMS = grammar, usage, mechanics, and style.

Table B15

Correlation Between Human Score/Grade and Normalized e-rater Microfeatures for e-rater Grammar Features (All Essay Orders)

	Persuasive essay human score	Descriptive essay human score	Grade level
GUMS 101 (sentence fragments)	-.05 *p* < .24	-.03 *p* < .55	-.05 *p* < .001
GUMS 103 (garbled sentences)	-.05 *p* < .25	-.05 *p* < .25	.00 *p* <59
GUMS 104 (subject-verb agreement)	-.03 *p* < .40	-.13 *p* < .001	.01 *p* < .42
GUMS 105 (ill-formed verb)	-.02 *p* < .67	.02 *p* < .63	.01 *p* < .13
GUMS 106 (pronoun error)	-.01 *p* < .72	n/a	-.02 *p* < .007
GUMS 107 (possessive error)	-.01 *p* < .75	.08 *p* < .001	.08 *p* < .001
GUMS 108 (wrong or missing word)	.00 *p* < .96	-.03 *p* < .43	.00 *P* < .64
GUMS 109 (proofread this)	-.08 *p* < .05	-.11 *p* < .007	-.01 *p* < .07

Note. GUMS = grammar, usage, mechanics, and style.

Appendix C

Existing and Proposed Scoring Rubrics for Cognitively Based Assessments of, for, and as Learning (CBAL)

Writing Strands I and II (as Used in the Pilot and Proposed Revision)

Fall Pilot, Generic Rubric: Strand I	
Score	Characteristics of a typical response (one or more paragraphs) at each score level
4 Excellent	An excellent response displays: Correct, well-formed sentences, varied in length and structure for effective communication A wide range of vocabulary, precise and well-chosen Few, if any, errors in grammar and mechanics, and spelling
3 Adequate	An adequate response displays: Reasonably well-formed sentences, varied in length and structure for clear communication A range of vocabulary, with words used appropriately Only minor errors in grammar, mechanics, and spelling—not serious enough to impede ease of reading
2 Limited	A limited response displays one or more of the following problems: Little variety in sentence length and structure Poor word choice, possibly causing some confusion Numerous errors in grammar and mechanics, which may occasionally impede ease of reading Numerous spelling errors, which occasionally impede ease of reading Uneven control of sentence structure/word order, which occasionally impedes coherence
1Minimal	A minimal response displays one or more of the following problems: No variety in sentence length and structure Frequent misuse of words or extremely limited word choice Serious and pervasive errors in grammar and mechanics, which frequently disrupt ease of reading Serious spelling errors, which frequently disrupt ease of reading Persistent lack of control over sentence structure, which frequently disrupts coherence
0 No Credit	A response receives no credit for any one of the following reasons: Not long enough for sentence-level characteristics to be judged Not written in English Off topic Blank Random keystrokes

Fall Pilot, Generic Rubric, Strand II	
Score	Characteristics of a typical response (one or more paragraphs) at each score level
4 Excellent	An excellent response: Has an effective overall structure, with content organized logically throughout Is well focused and coherent, with a clear relationship between main and subordinate ideas and clear transitions and connections between ideas Is well developed; main ideas are substantially supported with reasons, examples, facts, or other types of elaboration
3 Adequate	An adequate response: Has a clear overall structure, with content organized logically throughout most of the composition Is focused and coherent, with a generally clear relationship of main/subordinate ideas and appropriate transitions and connections between ideas Is adequately developed; main ideas are supported with reasons, examples, facts, or other types of elaboration
2 Limited	A limited response displays one or more of the following problems: Is poorly structured and organized Is poorly focused and weak in coherence Is insufficiently developed, with few reasons, examples, facts, or other types of support
1 Minimal	A minimal response displays one or more of the following problems: Is unstructured and disorganized throughout Lacks focus and coherence throughout Is undeveloped, lacking reasons, examples, facts, or other types of support
0 No Credit	A response receives no credit for any one of the following reasons: Not long enough for document-level characteristics to be judged Not written in English Off topic Blank Random keystrokes

Proposed Rubric for Strand IA, Master Academic Language	
Score	Characteristics of a typical response (one or more paragraphs) at each score level
4 Excellent	An excellent response displays: Effective choice of style and register to suit the task. In particular, an excellent response Effectively uses (but does not overuse or misuse) grammatical constructions typically associated with written style, such as passives, logical connectives, and attributive adjectives Where appropriate, effectively uses grammatical constructions typically associated with an oral or colloquial style A wide variety of sentence types and grammatical constructions, effectively chosen for clear communication. In particular, an excellent response Varies word choice and sentence structures effectively to maintain clarity and interest Consistently maintains clear reference and systematically avoids other forms of ambiguity Consistently avoids unnecessarily complex or confusing sentence patterns A wide range of vocabulary, precise and well chosen. In particular, it Uses all words accurately and idiomatically Uses topic-specific words (tier III words) to communicate clearly and precisely Uses typically academic words (tier II words) where appropriate Effectively uses (but does not overuse or misuse) abstract language and nominalizations Effectively uses more complex, Latinate vocabulary in a way that demonstrates mastery of a variety of word-building techniques Makes effective, vivid use of common, simple vocabulary (tier I words)

Proposed Rubric for Strand IA, Master Academic Language	
Score	Characteristics of a typical response (one or more paragraphs) at each score level
3 Adequate	An adequate response displays: A reasonable ability to adopt a style and register appropriate to the task. In particular, an adequate response Shows reasonable control of written style, making some use of typically written constructions such as passives, logical connectives, and attributive adjectives Shows reasonable control of spoken style, making use of typically oral/conversational patterns only when appropriate, and consistently avoiding inappropriate use of oral patterns in written contexts. Reasonably well-formed sentences, varied in length and structure for clear communication. In particular, an adequate response Varies word choice and sentence structures enough to avoid awkwardness and redundancy Generally maintains clarity of reference and avoids serious ambiguities in expression Contains relatively few sentences that are simultaneously complex and confusing A range of vocabulary, with most words used appropriately. For example, an adequate response Rarely misuses words. When errors in word choice appear, they only occur with infrequent, academically oriented or morphologically complex vocabulary Uses some appropriate topic-specific vocabulary (tier III words) Makes reasonable use of at least relatively common abstract, academic, Latinate words (tier II words) without lapsing into an awkward, obscure style Shows reasonable ability to choose clear phrasing using ordinary vocabulary (tier I words) without lapsing into an overly repetitive or oral style
2 Limited	A limited response displays one or more of the following problems: Occasional lapses into an inappropriately oral style, characterized by a lack of syntactic variety, a vocabulary consisting almost entirely of ordinary everyday words, an overuse of pronouns, and/or a subjective presentation focused on inappropriate expressions of personal opinion and reaction Occasional lapses into an awkward, unclear style, characterized by repetitive word choice, unclear references, uneven control of sentence structure, and/or ambiguity in expression Inappropriate word choices, possibly reflecting an attempt to use vocabulary above the students' normal level of expressive mastery

Proposed Rubric for Strand IA, Master Academic Language	
Score	Characteristics of a typical response (one or more paragraphs) at each score level
1 Minimal	A minimal response displays <u>one or more</u> of the following problems: An excessively oral style, characterized by a lack of syntactic variety, a vocabulary consisting almost entirely of ordinary everyday words, an overuse of pronouns, and/or a tendency toward a subjective presentation in which expressions of personal opinion and reaction are excessively common A persistently awkward, unclear style, characterized by repetitiveness, unclear references, uneven control of sentence structure, and/or ambiguity in expression Frequent, inaccurate word choices likely to cause confusion, involving vocabulary critical for the task
0 No Credit	A response receives no credit for any one of the following reasons: Not enough of the student’s own writing for sentence-level characteristics to be judged Not written in English Off topic Blank Random keystrokes

Proposed Rubric for Strand IB (Follow Written Convention)	
Score	Characteristics of a typical response (one or more paragraphs) at each score level
4 Excellent	An excellent response displays: Few, if any errors in grammar Few, if any errors in spelling Few, if any errors in capitalization and punctuation; capitalization and punctuation patterns are both standard and consistent Few errors reflecting hasty, careless text production, such as missing or repeated or inverted words Few, if any errors in usage for common grammatical categories (articles, prepositions, etc.)
3 Adequate	An adequate response displays: Only minor errors in grammar. Those grammar errors that do appear are the sort that are in fact entirely appropriate in informal or oral contexts for Standard English, but are normally edited out from formal, written documents. Only minor errors in spelling. Most of the spelling errors that do appear are of the sorts that reflect typographical errors rather than a lack of knowledge of how to transpose words into standard written form. There may be some errors that reflect lack of spelling knowledge for rarer words, but they should still be reasonable spellings in terms of the underlying orthographic systems of the language, with little or no confusion of common homophones. Few or no errors in the capitalization and punctuation necessary to indicate basic clause structure (including few or no run-ons or comma splices). Other capitalization or punctuation errors may be present as long as they do not impede ease of reading. Careless errors such as missing, repeated, inverted words are not so frequent as to impede ease of reading Usage of articles, prepositions, and other common grammatical categories is generally correct and where incorrect, seldom impedes ease of reading

<table>
<tr><th colspan="2">Proposed Rubric for Strand IB (Follow Written Convention)</th></tr>
<tr><th>Score</th><th>Characteristics of a typical response (one or more paragraphs) at each score level</th></tr>
<tr><td>2
Limited</td><td>A limited response displays one or more of the following problems:
Numerous errors in grammar, which may occasionally impede ease of reading
Numerous spelling errors, reflecting incomplete mastery of English orthography. Misspellings tend to be plausible rather than confused (for instance, misspellings may often involve the wrong vowel spelling for an unstressed syllable). This may also involve frequent confusion of common homophones.
Some errors in capitalization and punctuation that confuse basic clause structure, or else frequent minor errors in capitalization and punctuation, such as misuse of apostrophes and hyphens.
A significant number of careless errors, such as typos and instances where words are missing, repeated or inverted, sufficient to impede ease of reading.
Errors in the usage of articles, prepositions and other common grammatical categories, sufficient to impede ease of reading</td></tr>
<tr><td>1
Minimal</td><td>A minimal response displays one or more of the following problems:
Serious and pervasive ungrammaticality of the text produced
Serious spelling errors reflecting minimal knowledge of standard English orthography, including pervasive confusion of common homophones
Consistent failure to follow fundamental conventions for punctuation and capitalization, such as a consistent failure to mark sentence boundaries with appropriate use of periods and capitals, or else pervasive minor errors in capitalization and punctuation
A pervasive pattern of text production errors, such as typos or instances where words are missing, repeated, inverted, or otherwise confused in ways that disrupt ease of reading.
Systematic errors in usage</td></tr>
<tr><td>0
No Credit</td><td>A response receives no credit for any one of the following reasons:
Not enough of the student's own writing for sentence-level characteristics to be judged
Not written in English
Off topic
Blank
Random keystrokes</td></tr>
</table>

Proposed Revised Rubric for Strand II (Plan/Structure Documents)	
Score	Characteristics of a typical response (one or more paragraphs) at each score level
4 Excellent	An excellent response: Has an effective overall structure, with content organized logically throughout. There are clear paragraph breaks; each paragraph has a clear purpose and the overall structure and purpose of the document is easily grasped. The author consistently avoids tangents, and when elaborates on subsidiary ideas, clearly indicates how they fit into the overall document structure. This may or may not be achieved by explicit thesis and topic sentences, use of explicit transition words, placement of headings, and other explicit methods of indicating overall document structure, but the main idea of the document and of its component parts are clearly indicated and easily understood. [Select/Organize] Is well focused and coherent, with clear indications of the relative importance of ideas and easily grasped links and transitions between ideas . The writer makes effective use of complex sentences structures to indicate relative importance and logical relationships, and structures both sentences and sentence sequences so that the reader is easily able to follow the flow of ideas. [Focus/Connect] Is well developed; main ideas are substantially supported with reasons, examples, facts, or other types of elaboration [Activate/Retrieve]
3 Adequate	An adequate response: Has a clear overall structure, with content organized logically throughout most of the composition. Where paragraph breaks are missed, other textual cues make the shift in topic clear. There may be occasional tangents or apparently irrelevant material, but not so long or so frequent as to detract from the essential unity of the composition. [Select/Organize] Is focused and coherent, with a generally clear indication of how one idea relates to the next. There may be places where the connection between ideas is unclear and has to be inferred, but the gaps and jumps are not so large or so problematic as to obscure the general intent of the text. The author makes some use of complex sentence structures and other syntactic devices to integrate and focus ideas across clauses. [Focus/Connect] Is adequately developed; main ideas are supported with reasons, examples, facts, or other types of elaboration. [Activate/Retrieve]

Proposed Revised Rubric for Strand II (Plan/Structure Documents)

Score	Characteristics of a typical response (one or more paragraphs) at each score level
2 Limited	A limited response displays one or more of the following problems: Is poorly structured and organized. This may involve frequent tangents, lack of an overall outline (so that the discussion wanders randomly from one point to the next), and the like. There may be no clear overall thesis, and even if there is, it may be hard to identify the topics for individual paragraphs. [Select/Organize] Is poorly focused and weak in coherence. Transitions between ideas may not be explicitly marked even where the reason for the transition is hard to infer. Seemingly unimportant ideas may be emphasized by the way sentences are phrased, and it may take some work for the reader to follow the flow of ideas. [Focus/Connect] Is insufficiently developed, with few reasons, examples, facts, or other types of support [Activate/Retrieve]
1 Minimal	A minimal response displays one or more of the following problems: Lacks structure and organization. This may take the form of a short response without paragraph structure when the task requires an elaborate document, or it may take the form of a rambling, 'stream-of-consciousness text' at greater length that is unstructured and disorganized throughout. It will be difficult to identify the main idea or subsidiary topics with any certainty. Lacks focus and coherence throughout. Topic shifts are unpredictable and illogical. It may consistently use short, choppy sentences in an additive way, with no formal indication of how ideas are related across clauses. Is undeveloped, lacking reasons, examples, facts, or other types of support
0 No Credit	A response receives no credit for any one of the following reasons: Not enough of the student's own writing for document-level characteristics to be judged Not written in English Off topic Blank Random keystrokes