Listening. Learning. Leading.®

Research Report
ETS RR–14-03

Using Writing Process and Product Features to Assess Writing Quality and Explore How Those Features Relate to Other Literacy Tasks

Paul Deane

June 2014

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Using Writing Process and Product Features to Assess Writing Quality and Explore How Those Features Relate to Other Literacy Tasks

Paul Deane

Educational Testing Service, Princeton, NJ

This paper explores automated methods for measuring features of student writing and determining their relationship to writing quality and other features of literacy, such as reading rest scores. In particular, it uses the *e-rater*® automatic essay scoring system to measure *product* features (measurable traits of the final written text) and features extracted from keystroke logs to measure process features (measurable features of the writing process). These techniques are applied to student essays written during large-scale pilot administrations of writing assessments developed for ETS's *CBAL*™ research initiative. The design makes it possible to explore the factor structures of these product and process features and to examine how well they generalize beyond a single test session to predict underlying traits such as writing ability and reading level. Three product factors are identified, connected to fluency, accuracy, and content. The process factors are identified, corresponding to hesitancy behaviors, editing behaviors, and burst span (the extent to which text is produced in long bursts with only short internal pauses). The results suggest that writing process and product features have stable factor structures that generalize to predict writing quality, though there are some genre- or task-specific differences.

Keywords Writing; reading; assessment; keystroke log; e-rater; automated essay scoring

doi:10.1002/ets2.12002

A special issue of the *Journal on Writing Research* focused on the potential for using corpus data to study writing. Two of the articles in that special issue focused specifically on writing quality. Parr (2010) analyzed data from a large corpus containing more than 20,000 samples of student writing for 60 different prompts, focusing on developmental changes in writing quality scores across genres (conceptualized as *purposes for writing*). She found that progress across genres is uneven, with some purposes for writing (narration and reporting) eliciting much higher quality responses than others. Deane and Quinlan (2010) examined how well writing quality could be predicted using natural language processing (NLP) analysis techniques and analysis of keystroke logging data. Both studies contributed important elements: Parr (2010) contributed an examination of the differences in student performance across genres, whereas Deane and Quinlan (2010) explored the possibility of an analysis that automatically measures particular predictors of writing quality (such as organization, development, grammar, usage, spelling, mechanics, and typing speed) over an entire corpus of student work. This study examines the ways from where evidence can be gathered about both goals—elucidating variation across different occasions for writing and examining the extent to which automatically measured features can provide information about variations in writing quality—and also considers the relationship between writing quality and other literacy skills.

Many different factors can affect the level of writing quality achieved on any particular occasion, not all of them writing-specific, because there are strong links between reading, writing, and other literacy skills that can change over the course of development (Berninger, 1994). For instance, Berninger, Abbott, Abbott, Graham, and Richards (2002) found that reading comprehension exerts a direct, significant influence on compositional quality across the primary grades (1–6) and that other features (including spelling and compositional fluency) have significant covariance with compositional quality. Shanahan (2006), summarizing the previous literature on reading and writing, identified several major dimensions that account for these kinds of relationships:

1. shared orthographic, phonemic, lexical, syntactic, and discourse representations directly activated in both reading and writing.

Corresponding author: P. Deane, E-mail: pdeane@ets.org

2. more generally, a common substrate of shared abilities and representations, including domain knowledge; meta-knowledge about written language; and the visual, phonological, and semantic subsystems of language.
3. constraints on variance that arise from reading and writing interactions in normal development, in which, for instance, (re)reading may help to support development of planning and revision skills.

Constraints on variance that may arise from more general constraints, such as working memory span, also may yield connections between reading and writing because both reading and writing involve interactions between working memory and the internalization of literacy processes, with concomitant increases in fluency (McCutchen, Teske, & Bankson, 2008; Torrance & Galbraith, 2006). Given these considerations, it is important to establish methods to profile patterns of student performance and to examine the extent to which features characteristic of quality writing covary with other literacy skills, including reading comprehension. Differences among writers, and differences among genres, almost certainly reflect a complex array of factors that may partly reflect general levels of linguistic development and the general acquisition of literacy skills, while also reflecting specific writing strategies.

The methods advocated in Deane and Quinlan (2010) provide a way to provide detailed information about both the process and the product features of large corpora of electronically collected writing samples. These NLP features can provide information about the structure and development of a text, its content, its linguistic properties, and its adherence or nonadherence to conventions used in edited, published text. Analysis of keystroke logging features can provide information about the kinds of behaviors in which writers engage during the writing process, providing evidence not only about typing speed and accuracy but also about the prevalence of planning and editing behaviors and about how these behaviors are interleaved with text production. It is clear that many of these features are correlated with writing quality. However, there is as yet relatively little information about the kind or quality of the information that these automatically detected product and process features provide, and even less about how they relate to one another, to writing quality, or to other literacy processes such as reading. This paper will explore these questions by examining data from a large study in which reading and writing data were collected from more than 3,500 middle-school students in the United States.

Research Questions

The purpose of this study was exploratory: to determine how much information about writing quality (and related literacy skills) can be obtained using automatically extracted product and process features. This purpose can thus be distinguished from conventional approaches to automated essay scoring, where automated analysis is focused on predicting human scores for the same items for which features are extracted. Instead, this study is focused on determining how effectively automatically scored features could be used to generalize from individual performances to infer latent properties of the writer.

Because individual features may be unstable, they are first aggregated by identifying common factors. Relationships are examined between the resulting product and process factors and various criterion variables, including human essay scores, scores on selected-response and short constructed-response questions that are built into a lead-in or preparatory section in each writing test, and scores on separately administered reading tests. The following preliminary research questions are explored:

1. Does a stable factor structure emerge for product and process data across prompts?
2. Is there a relationship between the process and product factors?

The remaining questions are posed to explore how much information product and process factors provide and determine how far they generalize across multiple writing prompts or between even more distant, but related, literacy tasks:

3. Do both product and process factors contribute variance to the prediction of writing quality? Which of these factors are predictive of writing quality scores?
4. Do the product and process factors generalize? That is, do they predict scores across writing prompts? Across different asks (both within and across test forms)? Across both reading and writing tasks? Do we see evidence that specific factors vary by genre or topic or provide information primarily about writing, not reading?
5. Do the process and product features capture unique variance above and beyond that captured by human writing quality scores?

Table 1 Test Forms Administered

Sequence	First administration	Second administration
1	Ban Ads	Mango Street
2	Ban Ads	Wind Power
3	Ban Ads	Seasons
4	Mango Street	Ban Ads
5	Mango Street	Wind Power
6	Mango Street	Seasons
7	Wind Power	Ban Ads
8	Wind Power	Mango Street
9	Wind Power	Seasons
10	Seasons	Ban Ads
11	Seasons	Mango Street
12	Seasons	Wind Power

Methods

Instruments and Data Collection

The data analyzed in this study are drawn from a large multistate study in which each student took two reading, two writing, or a reading and writing assessment during spring 2011, in a counterbalanced design where students were assigned a pair of test combinations. Pairs of tests were distributed among schools, subject to the constraint that no test be administered on two different occasions in the same school. Tests assigned to a school were administered randomly within classrooms (Cline, 2012). A total of four assessments were administered, two reading and two writing, differing by genre, each named after the unifying theme or topic around which each test was designed. These consisted of the Ban Ads writing test form, focusing on persuasive writing; the Mango Street writing test form, focusing on literary analysis; the Wind Power reading test form, focusing on informational and persuasive texts; and the Seasons reading test form, focusing on literary texts. Each of these forms had been administered in prior pilot studies (see Bennett, 2011, for a summary of results). Detailed psychometric analyses of the tests and their relationships to each other are in preparation by other researchers. The design, in which each student took 1 of the 12 possible combinations of the four test sequences is shown in Table 1. The first and second administrations were spaced at least 3 months apart. This design allows comparison of two writing test forms that differ in genre (Sequences 1 and 4) of paired reading and writing tests (Sequences 2 – 3, 5 – 8, and 10 – 11) or of paired reading forms that differ in genre (Sequences 9 and 12). Each test form took two 45- to 50-minute class sessions to complete.

The assessments included in Table 1 are drawn from pilot assessments developed for ETS's *CBAL*™ (Cognitively Based Assessments *of*, *for*, and *as* Learning) research initiative (Bennett, 2011; Bennett & Gitomer, 2009). This initiative is part of an approach to K-12 assessment in the United States that combines several themes: multiple assessments throughout the school year, meaningful contexts, and rich tasks that can be viewed as learning experiences in their own right; these themes connect to a developmental framework that connects assessment explicitly with best practices in instruction and what is known about student learning and development from the cognitive and learning science literatures.

When this approach is applied to literacy skills — to reading and writing and their use in supporting thinking and learning — several major themes emerge (cf. Deane, 2011; Deane, Fowles, Baldwin, & Persky, 2011; Deane, Quinlan, & Kostin, 2011; Deane, Quinlan, Odendahl, Welsh, & Bivens-Tatum, 2008). In particular, one must recognize the importance not only of reading and writing but also of their coordination; and more specifically, one must recognize the importance of reading and writing being coordinated in meaningful contexts, where they define the literacy practices that students are expected to internalize. These themes emerge directly from the literature on the best practices in reading and writing instruction, where, for instance, Perin (2007) and Hillocks (2002) noted the importance of content and reasoning about content in the development of writing skill.

Previous Pilot Testing

The writing tests had been previously piloted as part of a larger set of writing assessments in fall 2009. Psychometric analyses indicated that these tests (which incorporated a series of prewriting tasks with a culminating essay) were generally

reliable (Cronbach's $\alpha > .7$). The reading tests had also been previously piloted (Bennett, 2011; Fu, Wise, & Chung, 2012), with high reliability (Cronbach's $\alpha = .89$ for one test form, .91 for the other) and also displayed correlations with state reading and/or state English language arts tests between .6 and .7. Dimensionality analyses revealed distinct reading and writing dimensions (Fu & Yoo, 2011; van Rijn, Deane, Rijmen, & Bennett, 2012; Yoo, Fu, Wise, & Chung, 2011) with some evidence of form-specific dimensions within each of those domains.

Design of the Writing Tests

The writing tests are designed to be taken in two class sessions. The two sessions are united by a common scenario and topic, such as analyzing a literary text or considering arguments on an issue. The first session consists of a series of lead-in tasks (a mixture of selected-response and constructed-response questions) intended to measure supporting skills and scaffold the students' understanding of, and engagement with, the topic and the content about which they would have to write in the second session. In particular, lead-in tasks require students' engagement with texts that provide content on the topic about which they will write and thus require them to engage in reading and thinking skills in preparation for writing on the same subject.[1] The second session is a single extended-writing task in which students are expected to produce a multiparagraph essay. The selection of lead-in tasks is governed by an analysis of the skills critical to a particular genre of writing; thus the Ban Ads design, which culminates in a persuasive essay, has lead-in tasks focusing on the analysis of argument and creating summaries from sources on the topic of "Should advertising to children under 12 years of age be banned?" Similarly, the Mango Street design, which culminates in a literary analysis of three passages from the novel *The House on Mango Street*, focuses on lead-in tasks in which readers are asked to make interpretive inferences and provide justifications for their interpretations.

The Lead-In Tasks: Ban Ads

The lead-in tasks that formed part of the Ban Ads test form comprised three short writing tasks and 21 selected-response questions. One block of five selected-response questions is excluded from analysis, as two of the five questions had problematic psychometric properties. The remaining questions were grouped into the following sets.

Summarization Tasks (Constructed Response)

Students were required to read two articles that were supplied to them as part of the test, both of which addressed the topic about which they were going to write, and then to prepare a short (two- to three-sentence) summary for each article.

Argument Classification Tasks (Selected Response)

Students were required to read 10 sentences, each of which implicitly took a position for banning or allowing advertisements to children under 12. They were then required to decide which side each sentence would be used to support.

Argument Evaluation Tasks (Selected Response)

Students read a sentence that explicitly took a position on the issue and provided an argument to support that position. They were then presented with a piece of evidence drawn from one of the articles and were asked to decide whether the evidence strengthened the argument, weakened it, or did neither.

Critique Task (Constructed Response)

Students read a letter to the editor presenting a straw man argument on the issue and were asked to write a paragraph-length critique of that argument, pointing out flaws in the reasoning.

The Lead-In Tasks: Mango Street

The lead-in tasks that formed part of the Mango Street form included a short constructed-response task and 10 selected-response questions. These included the following tasks.

Interpretive Comment (Constructed Response)

Write a two- to three-sentence response discussing the first chapter of Mango Street, after reading two other students' comments.

Identifying Textual Evidence (Five Selected-Response Questions)

Read five statements and find a sentence in the reading from Mango Street that supports each statement.

Evaluating Interpretations (Five Selected-Response Questions)

Choose the best global interpretive statements about all three readings from *The House on Mango Street*.

Details of the Reading Designs

The reading tests in this study are documented in Sheehan and O'Reilly (2011). These test forms, like the writing test forms, are designed to be taken in two class sessions, but they do not share a common scenario. Instead, one of the sessions focuses on a scenario, taking the form of a sequence of related reading tasks appropriate to the texts being read, and designed to sample across a range of skills and skill levels. The second session is devoted to blocks of unrelated reading tasks. While most of the questions are selected response, most of them use innovative formats that allow more natural interactions with text, such as selecting sentences in a passage or filling in a graphic organizer. For the purposes of analysis, therefore, the reading assessments include the Wind Power scenario block, which contains a series of passages and reading comprehension questions focused on informational texts about wind power, the Seasons scenario block, which contains a series of passages and reading comprehension items focused on literary texts about the seasons of the year, and two sets of independent (nonscenario) reading items, Block A and Block B. Each student who completed a reading test form was administered a scenario and an independent block in two class sessions that were scheduled as closely together as time permitted (within about 1 week).

Participants

Test forms were administered to 3,592 eighth-grade students attending 35 schools selected from 20 U.S. states. The schools were selected to sample from a variety of urban, suburban, and rural school districts representing a range of socioeconomic statuses and ethnicities. The sample was 49.5% male and 50.5% female. The ethnic composition was 63.5% White, 16.5% Hispanic, 11.8% Black, 7.7% Asian or Pacific Islander, and 0.3% Native American; 37.4% participated in free and reduced lunch programs. By English language classification, 85.8% were classified as proficient on school entry, 4.1% as former English language learners now English proficient, and 2.2% as English language learners; 6.9% were reported as qualifying for accommodation when taking Reading or Writing tests for annual yearly progress reporting.

Test Administrations

There was a 2-month gap between the two sessions required to administer the first test form and a second set of sessions required to administer the second form assigned to the same student. This gap was motivated by the need to make the administrations as realistic as possible for purposes of comparing performances under operational conditions.

Computerized tests were delivered over the Internet from an ETS server. Students took each test using their local schools' computing resources, either in a laboratory or in a classroom, depending on local arrangements. Test forms were administered in a counterbalanced partial block design in which each student took two of the four test forms in a randomized order. Final responses were collected, including all essays and short constructed-response answers, and a record of the amount of time spent on each item was made. In addition, keystroke logging data were collected for each essay, providing a detailed record of the timing of text production while essays were being composed by the students.

EXEMPLARY (5)
An EXEMPLARY response meets *all* of the requirements for a score of 4 *and distinguishes itself* with such qualities as insightful analysis (recognizing the limits of an argument, identifying possible assumptions and implications of a particular position); intelligent use of claims and evidence to develop a strong argument (including particularly well-chosen examples or a careful rebuttal of opposing points of view); or skillful use of rhetorical devices, phrasing, voice and tone to engage the reader and thus make the argument more persuasive or compelling.

CLEARLY COMPETENT (4)
The response demonstrates a competent grasp of argument construction and the rhetorical demands of the task, by displaying all or most of the following characteristics:

Command of Argument Structure

- States a clear position on the issue
- Uses claims and evidence to build a case in support of that position
- May also consider and address obvious counterarguments

- Makes reasonable claims about the issue
- Supports claims by citing and explaining relevant reasons and/or examples
- Is generally accurate in its use of evidence

Awareness of audience

- Focuses primarily on content that is appropriate for the target audience
- Expresses ideas in a tone that is appropriate for the audience and purpose for writing

DEVELOPING HIGH (3)
While a response in this categorydisplays considerable competence, it differs from Clearly Competent responses in at least one important way, such as a *vague claim; somewhat unclear, limited, or inaccurate use of evidence; simplistic reasoning; or occasionally inappropriate content or tone for the audience.*

DEVELOPING LOW (2)
A response in this category *differs from* Developing High responses because it displays problems that seriously undermine the writer's argument, such as *a confusing claim, a seriously underdeveloped or unfocused argument, irrelevant or seriously misused evidence, an emphasis on opinions or unsupported generalizations rather than reasons and examples, or inappropriate content or tone throughout much of the response.*

MINIMAL (1)
A response in this category *differs from* Developing Low responses in that it displays little or no ability to construct an argument. For example, there may be *no claim, no relevant reasons and examples, no development of an argument, or little logical coherence throughout the response.*

NO CREDIT (0)
Completely off task, consists almost entirely of copied source material, random keystrokes, blank, etc.

Figure 1 Rubric for Ban Ads test form focusing on rhetorical effectiveness and quality of argumentation.

Scoring

All tests were scored at ETS. Selected-response items were scored automatically by computer. Constructed-response items were scored by professional ETS scorers using rubrics and sample benchmark items from previous pilot administrations. Essays were scored by one scorer, except for a subset of 20% randomly selected essays that were rated by two scorers for quality assurance and for determination of rater reliability. Each essay received two scores: the first using a rubric specifically focused on the skills that distinguished the selected genre and a second using a generic rubric designed to measure general writing qualities. For instance, the Ban Ads essay prompt was scored using the rubrics shown in Figures 1 and 2, and the Mango Street essay prompt was scored using one common rubric (Figure 2) and a genre-specific rubric, shown in Figure 3. The writing tests were weighted so that the contribution of items to the total test score reflected the amount of time devoted to them; thus the human scores on each of the two rubrics for the essay prompt accounted for about half of the score total on each writing test. The following scores were available as a result:

1. total scores for each essay, combining the scores of the first human rater on each rubric, providing an estimate of writing quality.
2. total scores on the lead-in section of each writing test.
3. total scores on the four reading sections administered.

The lead-in sections contained both selected-response and short constructed-response (writing) tasks. The selected-response tasks were machine scored. The constructed-response tasks were scored on content rubrics. Errors in grammar, usage, mechanics, and style were ignored.

EXEMPLARY (5)
An EXEMPLARY response meets all of the requirements for a score of 4 but distinguishes itself by skillful use of language, precise expression of ideas, effective sentence structure, and/or effective organization, which work together to control the flow of ideas and enhance the reader's ease of comprehension.

CLEARLY COMPETENT (4)
A CLEARLY COMPETENT response typically displays the following characteristics:

It is adequately structured.
- Overall, the response is clearly and appropriately organized for the task.
- Clusters of related ideas are grouped appropriately and divided into sections and paragraphs as needed.
- Transitions between groups of ideas are signaled appropriately.

It is coherent.
- Most new ideas are introduced appropriately.
- The sequence of sentences leads the reader from one idea to the next with few disorienting gaps or shifts in focus.
- Connections within and across sentences are made clear where needed by the use of pronouns, conjunctions, subordination, etc.

It is adequately phrased.
- Ideas are expressed clearly and concisely.
- Word choice demonstrates command of an adequate range of vocabulary.
- Sentences are varied appropriately in length and structure to control focus and emphasis.

It displays adequate control of Standard Written English
- Grammar and usage follow SWE conventions, but there may be minor errors.
- Spelling, punctuation, and capitalization follow SWE conventions, but there may be minor errors.

DEVELOPING HIGH (3)
A response in this category displays some competence but differs from Clearly Competent responses in at least one important way, including *limited development; inconsistencies in organization; failure to break paragraphs appropriately; occasional tangents; abrupt transitions, wordiness; occasionally unclear phrasing; little sentence variety; frequent and distracting errors in Standard Written English; or relies noticeably on language from the source material.*

DEVELOPING LOW (2)
A response in this category differs from Developing High responses because it displays serious problems such as marked underdevelopment; disjointed, list-like organization; paragraphs that proceed in an additive way without a clear overall focus; frequent lapses in cross-sentence coherence; unclear phrasing; excessively simple and repetitive sentence patterns; inaccurate word choices; errors in Standard Written English that often interfere with meaning; or relies substantially on language from the source material.

MINIMAL (1)
A response in this category differs from Developing Low responses because of serious failures such as extreme brevity; a fundamental lack of organization; confusing and often incoherent phrasing; little control of Standard Written English; or can barely develop or express ideas without relying on the source material.

NO CREDIT (0)
Not enough of the student's own writing for surface-level features to be judged, not written in English; completely off topic, blank, or random keystrokes.

Figure 2 Generic rubric focusing on print, verbal, and discourse features.

Extraction of Product Features

E-rater Analysis

ETS's *e-rater*® automated essay scoring system allows both the prediction of overall scores and the extraction of features intended to measure specific aspects of student writing (Attali, 2011; Attali & Burstein, 2009; Attali & Powers, 2008; Burstein & Chodorow, 2002, 2010; Rock, 2007). The following features were the focus of this analysis:

- *Number of discourse units*. This feature is calculated from the output of a program that seeks to identify discourse units in the student's responses, such as topic and thesis sentences, introductions and conclusions, or supporting points. This feature is a logarithmic measure of the number of distinct discourse units identified.
- *Length of discourse units*. This feature is also calculated from the output of the discourse structure program (mentioned previously) and provides a logarithmic measure of the average length of identified discourse units.
- *Absence of mechanics errors*. This feature provides a square-root transformation of the number of mechanics and spelling errors observed in the essay, scaled negatively so that large numbers of errors result in a smaller value.

EXEMPLARY (5)
An EXEMPLARY response meets *all* of the requirements for a score of 4 *and distinguishes itself* with such qualities as *insightful analysis; thoughtful evaluation of alternative interpretations; particularly well-chosen quotations, details, or other supporting evidence; skillful use of literary terms in discussing the texts; or perceptive comments about the author's use of language, perspective, setting, mood, or other literary techniques.*

CLEARLY COMPETENT (4)
A typical essay in this category *presents an understanding of the story that includes not only surface elements (such as sequence of events) but also appropriate inferences about characters, their motivations, perspectives, interactions and/or development.* More specifically, it:

Analyzes and interprets the texts with reasonable clarity and accuracy

- *Goes beyond summarization by advocating a specific interpretation (or alternative interpretations) of the story as a whole*
- *Justifies the interpretation(s) by using relevant quotations, details, or other evidence from all three texts*
- *Makes clear connections between the interpretation and supporting evidence from the texts*

Shows an awareness of audience

- *Presents ideas in a way that makes it easy for the reader to see that the interpretation is valid*
- *Expresses ideas in a tone that is appropriate for the intended reader*

DEVELOPING HIGH (3)
While a response in this category displays considerable competence, it *differs from* Clearly Competent responses in at least one important way, such as a *simplistic or limited interpretation of the story (e.g., mentioning the writing but ignoring its importance); an interpretation based on fewer than three texts but which deals with the significance of Esperanza's writing; limited or occasionally inaccurate use of evidence; somewhat unclear or undeveloped explanations; mostly a summary; content not well-suited to the audience; or an occasionally inappropriate tone.*

DEVELOPING LOW (2)
A response in this category *differs from* Developing High responses in at least one of the following ways: a somewhat confusing or seriously limited interpretation (e.g., based on two texts but which ignores the writing); an interpretation based only on the third text; some inaccurate or irrelevant evidence from the story; an emphasis on opinions or unsupported statements; a confusing explanation of how the evidence supports the interpretation; merely a summary; or an inappropriate tone throughout much of the response.

MINIMAL (1)
A response in this category *differs from* Developing Low responses in that it displays little or no ability to justify an interpretation of literary texts. For example, there may be *an unreasonable or inaccurate interpretation of the story's characters, their motivations, perspectives, interactions and/or development; use of only the first or second text; a serious lack of relevant or accurate references to the text; a poor summary; or little coherence throughout the response.*

OFF-TOPIC (0)
No ability to communicate relevant ideas without relying on source material, not written in English; completely off topic, blank, or random keystrokes.

Figure 3 Genre-specific rubric for literary analysis (applied to essays written to the Mango Street prompt).

- *Absence of grammar errors*. This feature provides a square-root transformation of the number of grammatical errors observed in the essay, scaled negatively so that large numbers of errors result in a smaller value.
- *Absence of usage errors*. This feature provides a square-root transformation of the number of usage errors observed in the essay, scaled negatively so that large numbers of errors result in a smaller value.
- *Absence of stylistic errors*. This feature provides a square-root transformation of the number of stylistic errors (primarily excessive repetition) observed in the essay, scaled negatively so that large numbers of errors result in a smaller value.
- *Word length*. This feature provides a measure of vocabulary complexity, measured as the square root of the number of characters in the average word length.
- *Median word frequency*. This feature provides a measure of vocabulary familiarity, the median frequency of the words in the essay (as measured on a logarithmic scale).
- *Normal patterns of preposition and collocation usage*. This feature provides a measure of the extent to which prepositions and collocations are used in normal, idiomatic patterns, as determined by the analysis of a large corpus of edited English texts (the SourceFinder corpus used internally at ETS to select passages for inclusion in tests under development, which totals about 440 million words drawn primarily from books, magazines, and journal articles).

- *Content* (*pattern cosine and value cosine features*). These features provide a measure of the appropriateness of vocabulary, that is, the extent to which the words used in the student essay match those typically used in high-scoring or low-scoring essays written to specific prompts.

Experimental Features

Certain experimental product features were also included in the analysis, as follows:

- *Syntactic variety*. In addition to standard e-rater features, described earlier, the grammatical patterns present in a text provide a great deal of information about its style, register, and genre. Skilled writers make use of a variety of different syntactic patterns, whereas less-skilled writers may rely on a smaller range of sentence patterns. Automated analysis was applied to student responses to extract information about the rates of which more than 50 syntactic patterns appeared in student responses, including the presence of such signature written constructions as nominalizations and passives. We defined experimental features by conducting regressions to aggregate features measuring the rates at which these patterns appeared on a logarithmic scale. The aggregation models were trained against the scores for essays written for the *TOEFL*® and *GRE*® examinations, producing two different but related measures of syntactic variety.
- *Plagiarism analysis*. The writing tests required students to read specific informational or literary texts and then write about the same topic. This configuration, while encouraging combined reading and writing, also raises the risk that less able students will copy large chunks of text from the reading passages rather than putting information in their own words. Student responses were therefore analyzed using programs that identified chunks of text copied from one of the reading passages, and the proportion of the copied text in the overall response was calculated using a standard NLP technique.[2] Two prompt-specific features were created: one measuring the extent to which the text produced in an essay was original and one measuring the extent to which an essay included properly marked quotations.

Extraction of Process Features

The keystroke logging data were analyzed to identify several key event types previously identified from the writing literature (Almond, Deane, Quinlan, & Wagner, 2012) using a segmentation program that identified word and sentence boundaries from local cues, such as white space and punctuation marks. The analysis identified *bursts* (sequences of keystrokes without long pauses), cut/paste/jump events, backspacing events, and several types of pauses (between characters within a word, between words, and between sentences, among others). The analysis program calculated means and standard deviations for each event type. The discussion that follows examines these feature types in more detail and specifies how they were extracted from the keystroke log.

Bursts

The literature suggests that bursts — stretches of uninterrupted text production — provide a useful measure of the efficiency of writing (Chenoweth & Hayes, 2003). More skilled and fluent writers are able to produce text in relatively long bursts, reflecting an ability to plan and buffer text production more efficiently. As Almond et al. (2012) discussed, bursts were identified by breaking up the keystroke log at every pause longer than 2 standard deviations from the mean pause — about two thirds of a second. Burst length was measured by two features, both calculated from the logarithm of the number of characters in a burst: the mean and standard deviation of logarithmic burst lengths for each essay.

Within-Word Pauses

The literature suggests that latency between characters within a word will primarily be affected by the efficiency of typing and the presence or absence of pauses related to word finding and spelling behaviors (Torrance & Galbraith, 2006; Wengelin, 2007). The duration of within-word pauses formed a highly skewed distribution, which required measuring such pauses on a logarithmic scale. Once again, two summary features were extracted: the mean and standard deviation of logarithmic within-word pauses.

Between-Word Pauses

The literature suggests that variation in the latency between words (up to and including the first character in the next word) is primarily driven by word finding and editing processes, although some pauses between words reflect clause-boundary breaks, where planning processes are more likely (Wengelin, 2006). Pauses between words were also identified and measured on a logarithmic scale. As before, two summary features were extracted: the mean and standard deviation of logarithmic between-word pauses.

Between-Sentence Pauses

The literature suggests that pauses between sentences (up to and including latencies at the start of the initial words of the sentence) are associated with planning processes (Wengelin, 2006). Between-sentence pauses were also identified and measured on a logarithmic scale, with means and standard deviations calculated for each essay.

Single- and Multiple-Character Backspacing

The literature suggests that isolated backspace events are associated with spelling correction induced by self-monitoring during typing, whereas extended sequences of backspacing may be associated with editing behaviors, in which alternate words or longer phrasings are substituted on the fly when planning, text production, and phrase revision were interleaved (Chanquoy, 2009). To provide measurement relevant to such behaviors, summary statistics (means and standard deviations of logarithmic duration) were calculated for both single- and multiple-character backspacing events.

Cut/Paste/Jump Events

Editing actions that did not involve keystrokes — cutting, pasting, and jumping — presumably also provide direct evidence for the frequency with which a writer engages in editing behaviors. Once again, summary statistics (means and standard deviations of logarithmic durations) were calculated over all events of this type.

Expected Patterns

It was hypothesized that more proficient writers would display shorter mean durations for most event types and more consistent pause lengths at the level of individual keystrokes (reflecting faster and more efficient typing, spelling, and word finding) and would produce longer bursts of text production, showing greater variability in the length of pauses associated with planning, editing, and proofreading processes.

Data Analysis

Research Question 1

Product and process features were extracted for both prompts, and exploratory factor analysis (unweighted least squares with varimax rotation) was applied separately by prompt and feature type. It seemed plausible that the product features would display factor structures similar to those reported by Attali and Powers (2008), including factors for fluency, accuracy, vocabulary, and content. It also seemed plausible that the process features would display factor structures in which features measuring the speed of text production were differentiated from features more strongly associated with editing or planning behaviors, consistent with the literature on pauses during text production.

Research Question 2

Product – moment (Pearson's) correlation coefficients were calculated between the process and product factors associated with the same prompt. It was hypothesized that correlations between process and product factors within the same prompt would fall into the low to moderate range. If any strong correlations were observed, it would imply a need to postulate a common factor across the product and process features and would require running a single factor analysis instead of two separate factor analyses. Pearson's bivariate correlations were also calculated for the pairs of factors across prompts. The

highest correlations were expected to occur between the pairs of factors representing parallel feature sets, indicating that they represented the most similar constructs.

Research Questions 3–5

Each of the following questions can be addressed by conducting regression analyses in which the product and process factors are used to predict different criterion variables: scores on the other essay prompt, scores on lead-in tasks in either test form, and scores on the reading test forms. Because it will have been already established that each factor is no more than moderately related to any other, all predictor variables can be entered simultaneously. In each regression, a 20% cross-validation set is held back to help determine whether the models generalize to unseen data.

Research Question 3

To answer this question, regressions were performed predicting total human essay scores on the Ban Ads and Mango Street prompts from the product and process factor scores derived from automated analysis of responses to the same essay prompt. In each regression, two additional features were included, measuring the amount of original material not repeated from the source texts and the extent to which quotes were properly marked.

Research Question 4

Regressions were performed to predict total human essay scores for the Mango Street prompt from process and product factor scores derived from responses to the Ban Ads essay prompt. Similarly, regressions were performed to predict total human essay scores for the Ban Ads prompt from process and product factor scores derived from responses to the Mango Street essay prompt.

In addition, regressions were performed to predict scores on specific lead-in tasks from process and product factor scores derived from responses to both essay prompts.

Finally, regressions were performed to predict total scores on each of the four reading test forms from product and process factor scores derived from responses to both essay prompts.

In each regression, two additional features were included, measuring the amount of original material not repeated from the source texts and the extent to which quotes were properly marked.

Research Question 5

To address this question, partial correlations were calculated between all of the dependent and independent variables examined earlier in the study, controlling for human writing quality scores, and were compared to the pattern of regression weights obtained in the prior analyses. To the extent that the process and product factors capture unique variance above and beyond that associated with human writing quality scores, it can reasonably be concluded that they are providing additional information about student performances. Conclusions from the pattern of regression weights observed earlier in the study will be reinforced if the partial correlations follow the same patterns.

Results

Exploratory Factor Analyses and Correlations Between Product and Process Factors

Research Question 1

This question posits whether there is a stable factor structure emerging for process and product data across prompts.

Process Data

Roughly consistent with the previously cited literature, such as the confirmatory factor analyses in Attali and Powers (2008) and Attali (2011), the unweighted least squares exploratory factor analysis (with a Varimax rotation) yielded parallel, three-factor solutions.[3] These factors can be described as follows:

Table 2 Factor Weights for the First Three Factors for Product Features Derived From the 2011 Ban Ads and Mango Street Essay Responses

	Fluency		Accuracy		Content	
Feature	Ban Ads	Mango Street	Ban Ads	Mango Street	Ban Ads	Mango Street
Normal pattern of preposition and collocation usage	.68[a]	.63[a]	−.06	−.05	.05	−.11
Normal pattern of syntactic variety (GRE)	.75[a]	.70[a]	.11	.21	.08	−.10
Normal pattern of syntactic variety (TOEFL)	.56[a]	.53[a]	.09	.19	.00	−.19
Median word frequency	−.77[a]	−.79[a]	.63[b]	.58[b]	.01	.20
Average word length	.77[a]	.79[a]	−.50[b]	−.54[b]	.22	.01
Length of discourse unit	.78[a]	.75[a]	−.33	−.19	−.03	−.20
Number of discourse units	.71[a]	.52[a]	.29	.31	.01	−.15
Content (pattern cosine)	.66[a]	.75[a]	.03	.02	.73[c]	.44[c]
Content (value cosine, different aggregation)	−.62[a]	−.37[a]	.08	−.04	.55[c]	.91[c]
Absence of grammar errors	.05	.04	.84[b]	.82[b]	−.04	.02
Absence of usage errors	.07	.10	.82[b]	.85[b]	−.05	−.06
Absence of mechanics and spelling errors	.09	.13	.80[b]	.80[b]	.10	.03
Absence of stylistic errors	−.17	−.19	.61[b]	.63[b]	.06	−.07

Note: For Ban Ads, 73% of variance was explained by the first three factors; for Mango Street, 71% of variance was explained by the first three factors.

[a]Factors loadings >.35 on the fluency factor are highlighted in blue.

[b]Factor loadings >.35 on the accuracy factor are highlighted in orange.

[c]Factor loadings >.35 on the content factor are highlighted in purple.

- *Factor 1*: *Fluency*. Features that provide evidence that the writer can produce multiple-paragraph, elaborated text with generally appropriate content, a reasonable range of vocabulary words, and a distribution of grammatical constructions characteristic of fluent users of the English language.
- *Factor 2*: *Accuracy*. Features that indicate that the writer has difficulty going beyond relatively simple vocabulary or syntax without producing errors in conventions of grammar, usage, mechanics, and/or style.
- *Factor 3*: *Content*. Features that indicate whether the vocabulary used in the text is consistent with that used by higher or lower scoring writers on the selected prompt.

The consistency of the patterns of the feature weights observed across the two writing prompts suggests that there is, in fact, a stable cross-prompt factor structure for the product features examined in this study, though the features that reflect the use of the reading passages are an exception. Factor analyses that included these features yielded obvious asymmetries between the two prompts and an interaction between the content features and the features for use of the reading passages, and so those features were excluded from the factor analysis shown in Table 2. It would not be surprising if the pattern of plagiarism and quotation from sources differed between the two prompts, because in the case of the Ban Ads form, sources would primarily be cited to provide evidence backing up an argument, whereas the focus of the Mango Street form would naturally have led writers to quote from and discuss specific passages to support a literary interpretation.

In Table 2, the heaviest weightings on each factor are highlighted in contrasting colors. Note that unlike the factor analyses reported by Attali and Powers (2008) and Attali (2011), the NLP features associated with vocabulary were not borne out as a separate factor from fluency, accuracy and content, and loaded instead primarily on the fluency factor. It is not clear why this result was observed, although the population (eighth grade) is likely to show a smaller range of vocabulary knowledge than might be observed in an adult population.

Product Data

When an unweighted least squares exploratory factor analysis with a Varimax rotation was conducted, and all principal factors above an eigenvalue of 1 were extracted, a three-factor solution was obtained for both prompts, with closely parallel structure, as follows (see Table 3):

- *Factor 1*: *Latency*. This is a latency, or text production speed, factor, loading on the mean time spent in nearly every kind of keystroke event, and loading on the standard deviation of local pauses (between and within words).

Table 3 Factor Weights for the First Three Factors for Process Features Based on Keystroke Features Derived From the 2011 Ban Ads and Mango Street Essay Responses

	Latency		Editing behaviors		Burst span	
Feature	Ban Ads	Mango Street	Ban Ads	Mango Street	Ban Ads	Mango Street
Mean log burst length in characters	.19	.03	.15	.08	.76[c]	.76[c]
Mean log in-word pause	.89[a]	.89[a]	.08	.03	.13	.03
Mean log between-word pause	.90[a]	.89[a]	.12	.01	.20	.05
Mean log between-sentence pause	.56[a]	.49[a]	.41	.37	.31	.25
Mean log time spent insingle-character backspace	.77[a]	.67[a]	.26	.20	.25	.17
Mean log time spent inmultiple-character backspace	.67[a]	.67[a]	.36	.25	.29	.26
Mean log time spent in cut/paste/jump event	.36	.32	.44	.39	.18	.08
S.D. log burst length	.39	.30	.28	.20	.81[c]	.80[c]
S.D. log in-word pause	.64[a]	.66[a]	.16	.13	.09	.05
S.D. log between-word pause	.64[a]	.60[a]	.41	.35	.14	.09
S.D. log between-sentence pause	.21	.14	.56[b]	.53[b]	.31	.36
S.D. log time spent insingle-character backspace	.02	−.02	.57[b]	.64[b]	.24	.05
S.D. log time spent in multiple-character backspace	.33	.29	.57[b]	.52[b]	.18	.18
S.D. log time spent in cut/paste/jump event	.19	.16	.68[b]	.65[b]	.09	.05
Total time on task	.17	.14	.83[b]	.88[b]	−.04	−.05

Note: For Ban Ads, 73% of variance was explained by the first three factors; for Mango Street, 62% of variance was explained by the first three factors.

[a]Features loading >.5 on latency are highlighted in blue.

[b]Features loading >.5 on accuracy are highlighted in green.

[c]Features loading >.5 on burst span are highlighted in orange.

Table 4 Correlation Between Product and Process Factors for the Ban Ads Prompt

Factor	Latency	Editing	Burst span
Fluency	.08**	.58**	.34**
Accuracy	−.29**	.25**	.03
Content	−.02	.02	.01

Note: $N = 1{,}569$.

*$p < .05$.**$p < .001$.

- *Factor 2*: *Editing behaviors*. This is an editing factor, loading on the standard deviation of planning- and editing-related events, including between-sentence pauses, cut/paste/jump events, and backspacing. It also loads heavily on time on task.
- *Factor 3*: *Burst span*. This is a factor that loads almost entirely on the mean and standard deviations of burst length and thus on the writer's ability to produce relatively large chunks of text without stopping.

These factors can be given a relatively natural cognitive interpretation. Factor 1 is probably related to typing speed or at least to the general efficiency of transcription processes. Factor 2 is probably related to the general persistence at writing that leads to longer texts, more time on task, and editing behaviors that rework text, even if it means deleting or modifying already-produced parts of the text. Factor 3 is probably related to general verbal fluency or at least to those aspects of working memory that enable the writer to hold relatively large chunks of text in memory, that is, evidence of memory span for text production (burst lengths).

Research Question 2

This question asks whether there is a relationship between the process and product factors. The process factors have weak to moderate correlations with the product factors (see Tables 4 and 5). The latency factor has a weak negative relationship with the accuracy factor ($R = -.29$ for Ban Ads, $R = -.26$ for Mango Street). The editing factor is most strongly associated with the fluency factor ($R = .58$ for Ban Ads, $R = .52$ for Mango Street) and with the accuracy factor ($R = .25$ for Ban Ads,

Table 5 Correlation Between Product and Process Factors for the Mango Street Prompt

Factor	Latency	Editing	Burst span
Fluency	.07*	.52**	.35**
Accuracy	−.26**	.32**	−.02
Content	.13**	−.12**	.12**

Note: $N = 1{,}284$.
*$p < .05$. **$p < .001$.

Table 6 Correlation Between Product and Process Factors Across Prompts

	Mango Street fluency	Mango Street accuracy	Mango Street content	Mango Street latency	Mango Street editing	Mango Street burst span
Ban Ads fluency	.48[a] **	.17**	.07	−.04	.32**	.25**
Ban Ads accuracy	.20**	.26[a] **	.23**	.05	.17**	.20**
Ban Ads content	.12**	.12**	.19[a] **	−.11**	.08*	−.01
Ban Ads latency	.06	−.23**	−.06	.12[a] **	−.09*	−.18**
Ban Ads editing	.36**	.17**	.10*	.03	.40[a] **	.11**
Ban Ads burst span	.27**	.08	.06	−.16**	.07	.58[a] **

Note: $N = 721$.
[a]Correlations between equivalent factors across prompts are highlighted in gray.
*$p < .05$. **$p < .001$.

$R = .32$ for Mango Street). The burst span factor is most strongly associated with the fluency factor ($R = .34$ for Ban Ads, $R = .35$ for Mango Street). All other associations are weaker and can be neglected here, including the source usage features, which are only weakly associated with either process or product factors.

When correlations between Ban Ads and Mango Street product and process factors are examined (see Table 6), as expected, all but one of the highest positive cross-prompt correlations were for factors defined in terms of the same features, although these correlations are relatively low, ranging from .12 to .58.

Regression Analyses

All regressions were calculated by entering all variables simultaneously: the three product factors, the three process factors, and the two features for the use of sources. Twenty percent of the data were held out for cross-validation. In addition, the correlations between the independent variable and human writing quality scores were also calculated, for purposes of comparison. Results for Research Questions 3–5 are shown in Tables 7 and 8 and can be summarized as follows.

Research Question 3

This question asks whether product and process factors contribute variance to the prediction of writing quality and which of these factors are predictive of writing quality scores. Details of all the regressions reported here and in subsequent sections are shown in Tables 7 (for Ban Ads factors) and 8 (for Mango Street factors).

All three process factors (latency, editing, and burst span) are significant predictors of writing quality for both prompts, yielding models that yield moderately strong predictions of writing quality ($R^2 = .68$ for Ban Ads, $R^2 = .60$ for Mango Street). In these analyses, two of the three product factors (fluency and accuracy) are significant predictors of writing quality for both writing prompts (beta weight for fluency: .40 for Ban Ads, .23 for Mango Street; beta weight for accuracy: .35 for Ban Ads, .34 for Mango Street). However, the content product factor is a significant predictor for the Ban Ads prompt but not for the Mango Street prompt (beta weight for content: .06 for Ban Ads). The features for the use of sources (i.e., the reading passages) also contributed to the models. For the Ban Ads prompt, the proportion of copied text from the reading passages was a significant predictor of writing quality (beta weight = .15). For the Mango Street prompt, both the proportion of copied text (beta weight = .20) and the presence of correctly quoted source material (beta weight = .14) are significant predictors of writing quality.

Table 7 Regressions of Scores From Two Writing Prompts and Four Reading Forms on Product and Process Factors Derived From the 2011 Ban Ads Essay Features

	Beta weights for independent variables														
Dependent variable	Fluency factor	Accuracy factor	Content factor	Latency factor	Editing factor	Burst span factor	Proportion original language	Proportion properly marked quotes	*N*	Correlation with total human score for the Ban Ads essay	*R* (20% cross-validation set)	R^2	Adj.R^2	*SE* of the estimate
Human essay scores (Ban Ads)	.40[a] **	.35[a] **	.06[a] **	−.10[b] **	.25[c] **	.11[c] **	.15[d] **	.04	1,483	n/a	.82	.68	.67	3.53
Human essay scores (Mango Street)	.29[a] **	.20[a] **	.17[a] **	−.13[b] **	.14[c] **	.13[c] **	.06	.10[d] **	655	.57	.57	.36	.35	2.75
Summarytasks (Ban Ads)	.38[a] **	.21[a] **	.20[a] **	−.11[b] **	.10[c] **	.08[c] **	.02	−.02	1,230	.56	.55	.35	.34	1.19
Argument critique (Ban Ads)	.35[a] **	.21[a] **	.19[a] **	−.11[b] **	.13[c] **	.06[c] *	.03	.00	1,249	.56	.61	.33	.33	1.93
Interpretive comment (Mango Street)	.26[a] **	.17[a] **	.16[a] **	−.09[b] *	.08	.05	−.10	−.15[d] *	583	.41	.38	.20	.19	1.47
Argument classification (Ban Ads)	.36[a] **	.15[a] **	.16[a] **	−.06[b] *	.02	.01	.03	.00	1,250	.38	.45	.19	.18	1.76
Strength of evidence (Ban Ads)	.29[a] **	.18[a] **	.22[a] **	−.04	.06	.05	.00	.00	1,250	.41	.42	.22	.21	1.34
Identifying textual evidence (Mango Street)	.29[a] **	.20[a] **	.19[a] **	−.12[b] **	.06	.01	.13	.01	554	.45	.50	.23	.22	1.22
Evaluating interpretations (Mango Street)	.39[a] **	.24[a] **	.16[a] **	−.06	.04	.05	.09	.02	581	.51	.56	.30	.29	1.61
Wind Power reading scenario	.34[a] **	.18[a] **	.27[a] **	−.12[b] *	.14[c] *	−0.04	.04	.04	329	.60	.66	.37	.36	4.92
Seasons reading scenario	.39[a] **	.28[a] **	.26[a] **	−.06	−0.07	−0.01	.08	−.03	374	.47	.50	.34	.32	4.44
Reading Independent Block A	.46[a] **	.17[a] **	.27[a] **	−.14[b] *	−0.04	−.01	.12	.13	303	.60	.71	.38	.36	3.67
Reading Independent Block B	.39[a] **	.20[a] **	.26[a] **	−.14[b] **	−0.04	−.04	.13[d] *	.10	355	.43	0.54	0.32	0.3	3.46

[a]Significant predictions for the product factors are highlighted in blue.

[b]Significant predictions for the latency factor are highlighted in yellow.

[c]Significant predictions for the editing and burst span factors are highlighted in orange.

[d]Significant predictions for the quotation features are highlighted in gray.

$^{**}p < .001$. $^{*}p < .05$.

Table 8 Regressions of Scores From Two Writing Prompts and Four Reading Forms on Product and Process Factors Derived From the 2011 Mango Street Essay Features

	Beta weights for independent variables													
Dependent variable	Fluency factor	Accuracy factor	Content factor	Latency factor	Editing factor	Burst span factor	Proportion original language	Proportion properly marked quotes	*N*	Correlation with total human score for the Ban Ads essay	*R*(20% cross-validation set)	R^2	Adj. R^2	*SE* of the estimate
Human essay scores (Ban Ads)	.26[a] **	.21[a] **	.26[a] **	−.09[b] *	.18[c] **	.23[c] **	−.04	−0.10	534	.57	.65	0.44	0.43	4.85
Human essay scores (Mango Street)	.23[a] **	.34[a] **	−0.05[a]	−.11[b] **	.32[c] **	.09[c] **	.20[d] **	.14[d] **	1,159	n/a	.74	0.60	0.60	2.20
Summarytasks (Ban Ads)	.29[a] **	.26[a] **	.22[a] **	−.07	.02	.23[c] **	.00	.15[d] **	477	.48	.47	.34	.33	1.23
Argument critique (Ban Ads)	.31[a] **	.21[a] **	.25[a] **	−.08[b] *	.06	.19[c] **	.03	.14[d] **	486	.52	.55	.33	.32	2.04
Interpretive comment (Mango Street)	.30[a] **	.20[a] **	.04	−.06[b] *	.13[c] **	.12[c] **	.13[d] **	.05	999	.48	.48	.27	.27	1.39
Argument classification (Ban Ads)	.22[a] **	.22[a] **	.24[a] **	−.03	.02	.13[c] **	.07	.17[d] **	486	.38	.46	.22	.21	1.70
Strength of evidence (Ban Ads)	.19[a] **	.22[a] **	.24[a] **	−.08	.06	.10[c] *	−.04	.06	485	.40	.33	.19	.18	1.36
Identifying textual evidence (Mango Street)	.23[a] **	.24[a] **	.23[a] **	.03	.15[c] **	.12[c] **	.06	.14[d] **	970	.45	.50	.30	.30	1.18
Evaluating interpretations (Mango Street)	.32[a] **	.26[a] **	.14[a] **	.00	.17[c] **	.13[c] **	.12[d] **	.13[d] **	995	.54	.57	.39	.38	1.55
Wind Power reading scenario	.34[a] **	.31[a] **	.29[a] **	−.02	.15[c] *	−0.01	−0.05	−0.07	336	.52	.55	.38	.36	4.62
Seasons reading scenario	.44[a] **	.23[a] **	.21[a] **	−0.01	−0.13	.12[c] *	.21[d] *	.21[d] *	250	.61	.62	.50	.48	4.20
Reading Independent Block A	.35[a] **	.32[a] **	.27[a] **	−.03	−0.11	.13[c] *	−0.03	−0.09	310	.55	.58	.39	.37	3.85
Reading Independent Block B	.30[a] **	.21[a] **	.20[a] **	−.11	.24[c] **	−0.09	−0.06	.19[d] *	230	.54	.64	.50	.48	3.27

[a] Significant predictions for the product factors are highlighted in blue.

[b] Significant predictions for the latency factor are highlighted in yellow.

[c] Significant predictions for the editing and burst span factors are highlighted in orange.

[d] Significant predictionsfor the quotation features are highlighted in gray.

** $p < .001$. * $p < .05$.

The models performed well, at levels comparable to operational e-rater scoring models. The process factors and the plagiarism features all contributed, indicating that they are capturing aspects of the human scores not directly captured by the product factors alone.

Research Question 4

This question asks whether the product and process factors generalize. That is, do they predict scores across writing prompts? Across different asks (both within and across test forms)? Across both reading and writing tasks? Do we see evidence that specific factors vary by genre or topic or provide information primarily about writing, not reading?

Mango Street essay scores can be predicted from both Ban Ads product and process features with an R^2 of .36. Ban Ads scores can be predicted from Mango Street product and process factors with an R^2 of .44. Once again, see Tables 7 and 8. The beta weights for Ban Ads are as follows: Fluency = .29, Accuracy = .20, Content = .17, Latency = −.13, Editing = .14, and Burst Span = .13. The beta weights for Mango Street are as follows: Fluency = .26, Accuracy = .21, Content = .26, Latency = −.09, Editing = .18, and Burst Span = .23. The Ban Ads feature for properly marked quotes also predicts Mango Street scores ($\beta = .10$). Otherwise, the source usage features do not function as significant predictors across prompts. While the models explain less variance than for the prompts from which they were generated (with R^2 of .36 instead of .68, and .44 instead of .60), reductions in R^2 are to be expected, especially because the prompts were administered 2 – 3 months apart, address different topics, and respond to different genre expectations.

Similarly, models predicting lead-in scores from product and process factors function quite well, even when generalizing across forms.

Ban Ads lead-in scores can be predicted from the Ban Ads factors and from the Mango Street factors (adjusted R^2 ranging from .18 to .34). Mango Street lead-in scores can be predicted from the Mango Street factors (adjusted R^2 ranging from .19 to .29) and from the Ban Ads factors (adjusted R^2 ranging from .27 to .38). Again, see Tables 7 and 8.

Within these models, the product and process factors display different patterns:

1. All three product factors (fluency, accuracy, and content) are significant correlates of performance on all but one of the lead-in tasks both within and across test forms. Fluency beta weights range from .19 to .40; Accuracy from .15 to .26; Content beta from .14 to .25.
2. The Ban Ads latency factor is a significant correlate of five of the seven lead-in tasks, including all the Ban Ads lead-in tasks, with beta weights ranging between −.09 and −.11. By contrast, the Ban Ads editing and burst span factors are significant predictors only for the two Ban Ads lead-in tasks that involve writing, with beta weights of .10 and .13 for the editing factor and .08 and .06 for the burst span factor.
3. The Mango Street burst span factor is a significant predictor in all seven lead-in tasks, with beta weights ranging from .10 to .23. By contrast, the Mango Street editing factor is a significant predictor only for the Mango Street lead-in tasks, with beta weights ranging from .13 to .17, and the Mango Street latency factor is significant only for the Ban Ads critique task ($\beta = -.08$).
4. Features for the use of sources behave differently across the two test forms. Only the proper quotation feature is significant for Ban Ads, and that only for the Mango Street interpretive comment task. By contrast, the Mango Street proper quotation feature is significant for all of the Mango Street lead-in tasks and for three of the four Ban Ads lead-in tasks, with beta weights ranging between .13 and .17. The Mango Street plagiarism feature is significant for two of the three Mango Street lead-in tasks (beta weights of .12 and .13) but for none of the Ban Ads lead-in tasks.

Models predicting reading test scores from product and process features also function rather well (again, see Tables 7 and 8). They yield correlations similar to the correlations between human essay scores and reading test scores. The R values obtained fall in the range .55 to .71 on the cross-validation data, not very different from the R^2 values observed when Ban Ads scores are predicted from Mango Street scores ($R = 61$ on the cross-validation set) or Mango Street scores from Ban Ads scores ($R = .65$ on cross-validation).

Within these models, the product and process factors display different patterns:

1. All three product factors (fluency, accuracy, and content) are significant predictors of score on all four reading test forms. Fluency beta weights fall in the range .30 to .46. Accuracy beta weights fall in the range .17 to .32. Content beta weights fall in the range .20 to .29.

Table 9 Partial Correlations (Factoring Out Essay Score on the Same Prompt) Between Various Literacy Tasks and the Product and Process Factors for the Ban Ads Essay

Dependent variable	Fluency factor	Accuracy factor	Content factor	Latency factor	Editing factor	Burst span factor	Proportion original language	Proportion properly marked quotes	*N*
Human essay scores (Mango Street)	.09[a] *	.09[a] *	.16[a] **	−.13[b] **	.01	.10[c] *	−.04	.10[d] *	642
Summary tasks (Ban Ads)	.12[a] **	.09[a] **	.21[a] **	−.10[b] **	.02	.06[c] *	−.05	.05	1,449
Argument critique (Ban Ads)	.09[a] **	.09[a] **	.20[a] **	−.11[b] **	.02	.06[c] *	−.05	.05	1,471
Interpretive comment (Mango Street)	.06	.05	.13[a] **	−.09[b] *	.00	.04	−.02	−.01	715
Argument classification (Ban Ads)	.09[a] **	.07[a] **	.16[a] **	−.04	−.03	−.01	−.02	.03	1,476
Strength of evidence (Ban Ads)	.07[a] **	.09[a] **	.23[a] **	−.04	.01	.03	−.04	.06	1,477
Identifying textual evidence (Mango Street)	.04	.12[a] **	.20[a] **	−.10[b] **	−.03	.02	.04	−.02	681
Evaluating interpretations (Mango Street)	.11[a] **	.11[a] **	.17[a] **	−.05	−.06	.03	.01	.05	713
Wind Power reading scenario	.06	.12[a] *	.28[a] **	−.11[b] *	.02	.08	−.11[d] *	.12[d] *	305
Seasons reading scenario	.15[a] **	.16[a] **	.29[a] **	−.07	.07	.03	.05	−.01	357
Reading Independent Block A	.15[a] *	.12[a] *	.21[a] **	−.15[b] **	−.02	.10	−.12[d] *	−.16[d] *	281
Reading Independent Block B	.13[a] *	.15[a] **	.32[a] **	−.11	.02	.01	.09	.07	339

[a]Significant correlations for the product factors are highlighted in blue.
[b]Significant correlations for the latency factors are highlighted in yellow.
[c]Significant correlations for the editing and burst span factors are highlighted in orange.
[d]Significant correlations for the quotation features are highlighted in gray.
$^{**}p < .001$. $^{*}p < .05$.

2. The Ban Ads latency factor is significant for three of the four reading test forms, with significant beta weights ranging between .12 and .14.
3. Two process factors derived from the Mango Street responses, the editing and burst span factors, are significant for two of the four reading test forms (editing significant βs: .15, .24; burst span significant βs: .12, .13).
4. Source usage features do not consistently function as significant predictors of reading above and beyond the prediction provided by the product and process factors, though the Mango Street proper quotation feature is a significant predictor for two of the four reading forms, with beta weights of .19 and .21.

These patterns are reminiscent of the patterns observed with the lead-in tasks, with the Ban Ads latency factor and the Mango Street editing and burst span factors generalizing to reading, along with the Mango Street proper quotation factor.

Research Question 5

This question asks whether the process and product features capture unique variance above and beyond that captured by human writing quality scores. When human writing quality scores are controlled for, the product and process factors still show a pattern of significant partial correlations that is very similar to the pattern of regression weights analyzed earlier in this article (see Tables 9 and 10):

1. The product factors display significant partial correlations toward nearly every independent variable examined, with the Mango Street interpretive comment task providing the only exceptions. These significant partial correlations range in magnitude from .07 to .41.
2. The Ban Ads latency factor generalizes fairly broadly, sharing added variance with nearly all writing tasks, Mango Street selected-response tasks, and two of the four reading tests. These significant partial correlations range in magnitude from .09 to .15.
3. The Mango Street burst span factor generalizes fairly broadly, sharing added variance with the other essay, most lead-in tasks, and two of the four reading forms, with significant partial correlations ranging between .09 and .22.
4. The Ban Ads burst span factor shares added variance only with other writing tasks. These significant partial correlations range in magnitude from .06 to .10. The Ban Ads editing factor has no significant partial correlations in these data.

Table 10 Partial Correlations (Factoring Out Essay Score) Between Various Literacy Tasks and the Product and Process Factors for the Mango Street Essay

Dependent variable	Fluency factor	Accuracy factor	Content factor	Latency factor	Editing factor	Burst span factor	Proportion original language	Proportion properly marked quotes	*N*
Human essay scores (Mango Street)	.29[a] **	.19[a] **	.21[a] **	−.13[b] **	.05	.21[c] **	−.10[d] *	.07	495
Summary tasks (Ban Ads)	.22[a] **	.12[a] **	.18[a] **	−.05	−.16[c] **	.20[c] **	−.05	.09[d] *	545
Argument critique (Ban Ads)	.18[a] **	.11[a] **	.24[a] **	−.07	−.11[c] **	.21[c] **	.00	.02	557
Interpretive comment (Mango Street)	.18[a] **	.10[a] **	.05	−.09[b] **	.00	.09[c] **	−.01	.03	1,152
Argument classification (Ban Ads)	.21[a] **	.09[a] *	.22[a] **	−.01	−.06	.11[c] **	−.01	.05	557
Strength of evidence (Ban Ads)	.14[a] **	.09[a] *	.18[a] **	−.04	−.11[c] *	.04	−.05	.02	556
Identifying textual evidence (Mango Street)	.18[a] **	.14[a] **	.21[a] **	.03	.01	.08[c] **	−.01	.05	1,118
Evaluating interpretations (Mango Street)	.27[a] **	.16[a] **	.13[a] **	−.02	.04	.12[c] **	−.02	.06	1,147
Wind Power reading scenario	.24[a] **	.16[a] **	.20[a] **	.01	.02	.02	.04	−.02	299
Seasons reading scenario	.31[a] **	.18[a] **	.25[a] **	−.03	.03	.16[c] *	.03	.04	235
Reading Independent Block A	.27[a] **	.18[a] **	.18[a] **	−.06	−.01	.11	.01	.03	275
Reading Independent Block B	.41[a] **	.28[a] **	.18[a] **	.03	.21[c] **	.17[c] **	−.08	.15[d] **	216

[a]Significant correlations for the product factors are highlighted in blue.
[b]Significant correlations for the latency factor are highlighted in yellow.
[c]Significant correlations for the editing and burst span factors are highlighted in orange.
[d]Significant correlations for the quotation features are highlighted in gray.
$^{**}p < .001$. $^{*}p < .05$.

5. The Mango Street latency factor shares added variance with the Ban Ads essay and the Mango Street interpretive comment task, both writing tasks. Otherwise, it has no significant correlations after human writing quality scores are factored out.
6. The Mango Street editing factor shares added variance primarily with the Ban Ads lead-in tasks, with significant partial correlations ranging from .11 to .16.
7. The features for use of sources diverge from the regressions, with relatively few significant correlations over and above the human scores. However, the Ban Ads plagiarism and proper quotation features share added variance with two of the four reading forms.

Discussion

Exploratory factor analysis appears to support the existence of consistent factor structures shared across prompts both for product and process features. In both writing prompts, the product features group into three factors with similar patterns of feature weights (fluency, accuracy, and content). In both prompts, the process features also group into factors with similar patterns of feature weights (latency, editing, and burst span). However, the relatively low correlations between the parallel factors across the two prompts suggest that students may respond differentially to the specific challenges posed by a particular prompt, especially where (as here) the prompts vary both in genre and topic. The exact parameters of this variation can only be ascertained by examining a much wider range of variations across genres and topics.

It is not particularly surprising that an argument essay and a literary analysis essay on entirely different topics should evoke very different patterns of student performance. It is well known that the genre and topic of a writing task can affect how well students will perform (Breland, Bridgeman, & Fowles, 1999; Breland, Lee, & Muraki, 2004; Hoetker & Brossell, 1989; Quellmalz, Capell, & Chou, 1982; Ruth & Murphy, 1988; Spaan, 1993). The relatively weak correlations suggest that product and process factors will be able to quantify differences in patterns of writer performance across prompts.

Conversely, the product factors derived from e-rater show a relatively stable relationship with a wide range of literacy skills. Although there is a significant fall-off in prediction when we try to predict scores across tasks or occasions, rather than using product features to characterize writing quality on the same prompt, the three product factors are significant predictors for almost every literacy task in the battery. Because they also have significant partial correlations with all of these literacy tasks after human essay scores are factored out, it seems safe to conclude, at the least, that they measure aspects of writing and associated literacy skills more reliably than the human scoring methods applied in this study. In some cases, the partial correlations are so large that increases in reliability cannot be the sole explanation for the partial correlations. The Mango Street fluency factor, for instance, shows partial correlations to reading tests between .24 and .41, which is consistent with the hypothesis that the measurement it provides is rather strongly related to some latent feature of reading skill, above and beyond its connection to writing.

Perhaps most interesting of all, the process factors appear to provide different information by prompt. A pattern of slow (high-latency) text production on the Ban Ads prompt appears to reflect a general pattern of lower performance on a wide range of literacy tasks. High-latency text production on the Mango Street prompt seems much more limited in its implications. Conversely, longer bursts of text production are predictive only of writing performance when measured using the Ban Ads prompt but seem to reflect more general features of literacy when measured using the Mango Street prompt.

It may be possible to find ways to measure process factors that can tease apart these kinds of differences among prompts. But even on an exploratory basis, the differences in patterns of performance are suggestive. For instance, we might consider the fact that the reading required for the Ban Ads form is relatively challenging. It is possible that relatively difficult texts could have an inhibitory effect on writing processes, which would account for the way that the Ban Ads latency feature generalizes to predict reading and thinking tasks, not just writing. The reading for Mango Street is relatively easy, which might account for less predictive value for the latency feature beyond text production processes. These suggestions are purely speculative, but they do suggest that the kinds of features we have extracted might be useful to support a detailed study of how prompt-by-task interactions affect student performance.

Conclusions

The results also suggest that the approach taken in this study will support a more general methodology for analyzing large writing corpora, by providing methods for automated analysis:

1. The product and process factors could be used to define profiles of writer performance, separating out groups that are high on one factor but low on another on particular types of prompts. If, as seems likely, differences on these dimensions are likely to correspond to major differences in skill and writing-strategy use, then it may be possible to use these factors to analyze dimensions of variation among writers drawn from the same population, including differences in the characteristics of writers fitting into different profiles and variations in the extent to which development along specific dimensions correlates with external measures such as reading scores.
2. To the extent that particular prompt types or writing genres turn out to load consistently more strongly on particular product or process factors, these in turn may indicate specific differences in style and processing between different kinds of writing.
3. If writing samples from student writers are collected across a developmental range, as Attali and Powers (2008) and Attali (2011) illustrate for e-rater features and factors derived from them, then changes in the predictive value of particular factors (and possibly changes in the factor structure over time) can provide a method to analyze the development of writing skill over large collections of student writing samples without having to manually evaluate the quality of every text in the corpus.

Given the specific population and design employed, there are obvious limitations to the conclusions that can be drawn from these results:

1. All the essay tasks assigned in this study required the writer also to read source texts. A different study would have to be conducted to disentangle the contribution of source-text reading from the contribution of more generic text production skills to student performances on this type of writing task.

2. Only two writing prompts were administered, differing both in genre and in topic. A different study would have to be conducted to disentangle the effects of genre from the effects of topic or to estimate how stably the product and process factors generalize across a range of prompt types and genres.

These limitations are consistent with the exploratory nature of this study.

This study can be seen as a preliminary example of this type of analysis, applied to demonstrate that automatically extracted features of student performances on a single writing prompt are strongly related to their general levels of reading and writing performance yet display differential patterns that may reflect differences in the cognitive processes and rhetorical strategies that writers apply to different kinds of writing tasks.

Acknowledgments

The work reported here has built on the contributions of current and former colleagues at ETS. The study analyzed here was administered by Fred Cline and Lauren Phelps. Kathleen Sheehan, Tenaha O'Reilly, and Heather Nadelman led the work that designed the reading forms. The writing test forms reflect the work (among others) of Douglas Baldwin, Jennifer Bivens-Tatum, Peter Cooper, Mike Ecker, Barbara Elkins, Mary Fowles, Nathan Lederer, Roseanne Morgan, Heather Nadelman, Norah Odendahl, Peggy Redman, Margaret Vezzu, Mike Wagner, and Cynthia Welsh. René Lawless helped with editing this paper. The analysis of the automated scoring and timing features reported here benefited from discussions and analyses conducted with Russell Almond, Yigal Attali, Irene Kostin, and Thomas Quinlan and from discussions with Randy Bennett, Peter van Rijn, John Sabatini, and Jianbin Fu. Various other colleagues have contributed in other ways, including Marjorie Biddle, Jennifer Caterson, and Lynn Zaback.

Notes

1 The lead-in section of a writing test may therefore include items that might be interpreted, in other contexts, as reading or critical-thinking items, but in context, they clearly function as prewriting tasks. Psychometric analyses indicate they do provide measurement of a common factor with the pure writing tasks, though there may also be genre-specific factors that distinguish the two writing tests when viewed at a finer grain size (Peter van Rijn, personal communication, April 2012).

2 The specific technique involved use of prefix tree representations derived from the source texts to identify segments of text in the candidate's writing that replicated long strings of characters from the source texts (but allowing for some small breaks or variations in the quoted sequence).

3 For Ban Ads, these three factors were identified by selecting factors with eigenvalues greater than 1. For Mango Street, the third factor fell below just below that threshold but upheld the structure shown when a three-factor solution was induced.

References

Almond, R., Deane, P., Quinlan, T., & Wagner, M. (2012). *A preliminary analysis of keystroke log data from a timed writing task* (Research Report No. RR-12-23). Princeton, NJ: Educational Testing Service.

Attali, Y. (2011). *Automated subscores for TOEFL iBT independent essays* (Research Report No. RR-11-39). Princeton, NJ: Educational Testing Service.

Attali, Y., & Burstein, J. (2009). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, *4*(3), 3–30.

Attali, Y., & Powers, D. (2008). *A developmental writing scale* (Research Report No. RR-08-19). Princeton, NJ: Educational Testing Service.

Bennett, R. E. (2011). *CBAL*: *Results from piloting innovative K–12 assessments* (Research Report No. RR-11-23). Princeton, NJ: Educational Testing Service.

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.

Berninger, V. W. (1994). *Reading and writing acquisition: A developmental neuropsychological perspective.* Madison, WI: Brown and Benchmark.

Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S., & Richards, T. (2002). Writing and reading: Connections between language by hand and language by eye. *Journal of Learning Disabilities*, *35*, 39–56. doi:10.1177/002221940203500104.

Breland, H., Bridgeman, B., & Fowles, M. (1999). *Writing assessment in admission to higher education*: *Review and framework* (Report No. 99–03). New York, NY: College Board.

Breland, H., Lee, Y.-W., & Muraki, E. (2004). *Comparability of the TOEFL CBT Writing Prompts*: *Response mode analyses* (TOEFL Research Report No. RR-75). Princeton, NJ: Educational Testing Service.

Burstein, J., & Chodorow, M. (2002). Directions in automated essay scoring. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 487–497). New York, NY: Oxford University Press.

Burstein, J., & Chodorow, M. (2010). Progress and new directions in technology for automated essay evaluation. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed., pp. 529–538). New York, NY: Oxford University Press.

Chanquoy, L. (2009). Revision processes. In R. Beard, D. Myhill, J. Riley, & M. Nystrand (Eds.), *The Sage handbook of writing development* (pp. 80–97). London, England: Sage.

Chenoweth, N. A., & Hayes, J. R. (2003). The inner voice in writing. *Written Communication*, *20*, 99–118.

Cline, F. (2012). *2011 CBAL multi-state reading and writing study*. Unpublished manuscript.

Deane, P. (2011). *Writing assessment and cognition* (Research Report No. RR-11-14). Princeton, NJ: Educational Testing Service.

Deane, P., Fowles, M., Baldwin, D., & Persky, H. (2011). *The CBAL summative writing assessment*: *A draft eighth-grade design* (Research Memorandum No. RM-11-01). Princeton, NJ: Educational Testing Service.

Deane, P., & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, *2*(2), 151–177.

Deane, P., Quinlan, T., & Kostin, I. (2011). *Automated scoring within a developmental*, *cognitive model of writing proficiency* (Research Report No. RR-11-16). Princeton, NJ: Educational Testing Service.

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* (Research Report No. RR-08-55). Princeton, NJ: Educational Testing Service.

Fu, J., Wise, M., & Chung, S. (2012). *Statistical report of fall 2009 CBAL reading tests* (Research Memorandum No. RM-12-12). Princeton, NJ: Educational Testing Service.

Fu, J., & Yoo, H. (2011, April). *Dimensionality analysis of Cognitively Based Assessment of*, *for*, *and as Learning* (*CBAL*) *grade 8 writing tests*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

Hoetker, J., & Brossell, G. (1989). The effects of systematic variations in essay topics on the writing performance of college freshmen. *College Composition and Communication*, *40*, 414–421.

McCutchen, D., Teske, P., & Bankson, C. (2008). Writing and cognition: Implications of the cognitive architecture for learning to write and writing to learn. In C. Bazerman (Ed.), *Handbook of research on writing: History, society, school, individual, text* (pp. 451–470). New York, NY: Lawrence Erlbaum Associates.

Parr, J. (2010). A dual purpose data base for research and diagnostic assessment of student writing. *Journal of Writing Research*, *2*(2), 129–150.

Perin, D. (2007). Best practices in teaching writing to adolescents. In S. Graham, C. MacArthur, & J. Fitzgerald (Eds.), *Best practices in writing instruction* (pp. 242–264). New York, NY: Guilford Press.

Quellmalz, E. S., Capell, F. J., & Chou, C. P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement*, *19*, 241–258.

Rock, J. L. (2007). *The impact of short-term use of Criterion*^{SM} *on writing skills in ninth grade* (Research Report No. RR-07-07). Princeton, NJ: Educational Testing Service.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex.

Shanahan, T. (2006). Relations among oral language, reading, and writing development. In C. A. MacArthur & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 171–186). New York, NY: Guilford Press.

Sheehan, K., & O'Reilly, T. (2011). *The CBAL reading assessment*: *An approach for balancing measurement and learning goals* (Research Report No. RR-11-21). Princeton, NJ: Educational Testing Service.

Spaan, M. (1993). The effect of prompt in essay examinations. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 98–122). Alexandria, VA: TESOL.

Torrance, M., & Galbraith, D. (2006). The processing demands of writing. In C. A. MacArthur & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 67–80). New York, NY: Guilford Press.

van Rijn, P. W., Deane, P., Rijmen, F., & Bennett, R. E. (2012, July). *Local dependence and dimensionality considerations in applying multidimensional IRT models to innovative writing assessment*. Paper presented at the 77th annual meeting of the Psychonomic Society, Lincoln, NE.

Wengelin, A. (2006). Examining pauses in writing: Theories, methods and empirical data. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer key-stroke logging and writing: Methods and applications* (Vol. *18*, pp. 107–130). Oxford, England: Elsevier.

Wengelin, A. (2007). The word level focus in text production by adults with reading and writing difficulties. In D. Galbraith, M. Torrance, & L. van Waes (Eds.), *Recent developments in writing process research. Volume 2: Basic processes* (pp. 67–82). Dordrecht, the Netherlands: Kluwer Academic Press.

Yoo, H., Fu, J., Wise, M., & Chung, S. (2011, April). *Dimensionality analysis of Cognitively Based Assessment of*, *for*, *and as Learning* (*CBAL*) *reading tests*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Action Editor: Beata Beigman Klebanov

Reviewers: Brent Bridgman and Chaitanya Ramineni