# Automated Essay Scoring in Innovative Assessments of Writing from Sources

**by Paul Deane, Frank Williams, Vincent Weng, Catherine S. Trapani**

## Abstract

This study examined automated essay scoring for experimental tests of writing from sources. These tests (part of the CBAL research initiative at ETS) embed writing tasks within a scenario in which students read and respond to sources. Two large-scale pilots are reported: One was administered in 2009, in which four writing assessments were piloted, and one was administered in 2011, in which two writing assessments and two reading assessments were administered. Two different rubrics were applied by human raters to each prompt: a general rubric intended to measure only those skills for which automated essay scoring provides relatively direct measurement, and a genre-specific rubric focusing on specific skills such as argumentation and literary analysis. An automated scoring engine (e-rater��®) was trained on part of the 2009 dataset, and cross-validated against the remaining 2009 dataset and all the 2011 data. The results indicated that automated scoring can achieve operationally acceptable levels of accuracy in this context. However, differentiation between the general rubric and the genre-specific rubric reinforces the need to achieve full construct coverage by supplementing automated scoring with additional sources of evidence.

## Introduction

### Goals and Motivation

The advent of the Common Core State Standards (CCSS) and the associated assessment consortia is likely to have a significant impact on the assessment of writing at the K-12 level in the United States. The CCSS, a major educational reform effort in the U.S., is designed to raise the standards for performance by linking successive levels by grade with the progress students need to be prepared for college and career training (Porter, McMaken, Hwang, & Yang, 2011). One change built into the CCSS is an increased emphasis on writing under conditions more like those needed for success in a college classroom. Under earlier assessment regimens, it was common for writing to be assessed using a single, timed writing prompt written without access to sources, focused on a general subject easily addressed by writers with a wide variety of backgrounds, knowledge, and experiences. This approach has been widely criticized on a variety of grounds (Murphy & Yancey, 2008), with particular emphasis from some critics on the risk that such assessments will encourage an approach to school writing that divorces it from meaningful, thoughtful engagement with content (Hillocks, 2002). Under the Common Core State Standards, much more emphasis is likely to be placed upon writing from sources. In fact, five of the ten college and career readiness anchor standards specified for writing emphasize engagement with content and source texts:

- *Standard 1*: Write arguments to support claims in an analysis of substantive topics or texts, using valid reasoning and relevant and sufficient evidence.
- *Standard 2*: Write informative/explanatory texts to examine and convey complex ideas and information clearly and accurately through the effective selection, organization, and analysis of content.
- *Standard 7*: Conduct short as well as more sustained research projects based on focused questions, demonstrating understanding of the subject under investigation.
- *Standard 8*: Gather relevant information from multiple print and digital sources, assess the credibility and accuracy of each source, and integrate the information while avoiding plagiarism.
- *Standard 9*: Draw evidence from literary or informational texts to support analysis, reflection, and research. (Common Core Standards Initiative, 2011)

These emphases are reflected, in turn, by specifications from both the consortia that emphasize writing from sources in the design of tests being created under aegis of the Race to the Top Assessment grants program (State of Florida Dept. of Education, 2012, p. 35; Smarter Balanced Assessment Consortium, 2012a, p. 17).

At the same time, there is increased emphasis on use of automated essay scoring (AES) to support large scale testing, particularly in the context of the Common Core State Standards. This emphasis has led to large-scale evaluations of automated scoring systems (Attali, 2013; Ramineni & Williamson, 2013; Shermis & Hamner, 2012). The Smarter Balanced Assessment Consortium (Smarter Balanced) anticipates heavy use of automated scoring (including AES), as does the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium, although it appears to anticipate a larger role for human scoring. For essay tasks, this emphasis on automated scoring is potentially in conflict with the emphasis on critical thinking and writing from sources, to the extent that the rubrics for essay scoring emphasize such elements as "valid reasoning," "relevant and sufficient evidence," or conveying "complex ideas ... accurately." As the articles collected in Shermis & Burstein (2003) indicate, AES systems do not explicitly evaluate the validity of reasoning, the strength of evidence, or the accuracy of information. They rely instead on measures of such things as the structure and elaboration of student essays, the sophistication of vocabulary, or the number of errors in grammar, usage, mechanics or style. This is a point that has been raised at length by critics of AES, such as the collection of articles presented in

Ericsson & Haswell (2006).

While the features measured by AES systems may provide useful proxies for the broader conception of writing expressed in the Common Core State Standards, there is a clear area of risk. The Race to the Top Assessment Program consortia plan to require writing from sources. They seek to measure a rich writing construct that includes strength of argumentation and clarity and accuracy of explanation. At the same time, their plans indicate the intention to rely heavily on automated scoring technologies, which use features that directly measure a narrower construct focused on surface form. It is important to determine under what conditions such a program is likely to be successful.

As it happens, we have been exploring closely related questions as part of an ETS research initiative identified by the acronym CBAL ("Cognitively-Based Assessments of, for, and as Learning"). This approach, which predates the emergence of the Race to the Top Assessment program and the Smarter Balanced and PARCC consortia, is intended to address known problems with K-12 assessment in the U.S. (Bennett, 2011; Bennett & Gitomer, 2009), by designing assessments to maximize the potential for positive effects on teaching and learning.

When the CBAL approach is applied to literacy skills - to reading, writing, and their use to support 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, we are forced to recognize the importance not only of reading and writing, but of their coordination; and more specifically, we must admit 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 (2009) and Hillocks (2002, 2008) note the importance of content and reasoning about content in the development of writing skill. And most importantly in the present context, they lead to test designs generally compatible with those proposed by the Smarter Balanced and PARCC consortia, in particular, to writing assessment designs that emphasize reasoning skills and writing from sources.

In our work on writing assessment under the CBAL initiative, we have considered key issues about how to measure higher-order reasoning and content skills while making effective use of AES technologies. We have piloted, scored, and built automated scoring models for a number of different assessment designs. These designs, and the results of the large-scale pilots, provide useful information that can be leveraged to achieve the major goal of this paper: to evaluate how effectively AES can be used when it is applied to innovative writing tests that focus on writing from sources.

**Strategies for Providing a Rich Measure of the Writing Construct**

Before considering the specific study presented in this paper, it will be useful to take a step back to consider, first, how one might assess the writing construct, and second, how automated essay scoring might combine with other sources of evidence. Writing is a complex construct, involving strategic coordination of disparate skills to achieve an integrated performance. This has implications for both learning and assessment. The literature suggests that stronger writers will have stronger component skills and better control of cognitive problem-solving strategies, both of which facilitate performance (Bereiter & Scardamalia, 1987; McCutchen, 2006). Thus, strong students tend to be strong across the board, while weak students may experience cascading failures in which problems in one area block progress in another. Perhaps as a result, trait scores for writing tend to be strongly correlated with one another and with holistic writing scores (Freedman, 1984; Huot, 1990; Lee, Gentile, & Kantor, 2008).

The complexity of writing as an integrated performance is an even greater issue when a writing task requires use of source materials. In such a situation, reading is inseparable from writing, and poor performance could derive from a variety of causes: failure to understand the source materials; failure to think about them in task-appropriate ways; failure to use appropriate planning strategies to address the task; inadequate argumentation skills; difficulties in text production; or general deficits in working memory and executive control processes. If the only information we have about this complex process comes in the form of a holistic score for the final written response, it will be impossible to disambiguate the causes of poor performance. It may well be impossible to address all of the possible causes of low performance in a single assessment; but to the extent that the assessment design can build in some cross-checks, it may be possible to distinguish groups of students with different characteristic profiles of performance. To accomplish this goal, it would be helpful if the assessment design provides multiple sources of evidence, beyond human holistic scores, with which to characterize student performance.

These issues have heavily influenced the development of innovative test designs for the CBAL research initiative. We have explored several strategies for collecting richer evidence about student performance on tasks that require writing from sources. Each strategy targets different aspects of the issues discussed above. For instance, items focused on student ability to handle argument critically may help disambiguate whether poor student performance is related to difficulties with interpreting argument. Similarly, information about the writing process, such as may be collected from a keystroke log, may help to disambiguate whether students are having difficulties in text production. We hypothesize that we will obtain better information about levels of student performance by applying several of these strategies in combination. These strategies are described briefly below.

***Lead-in tasks.***

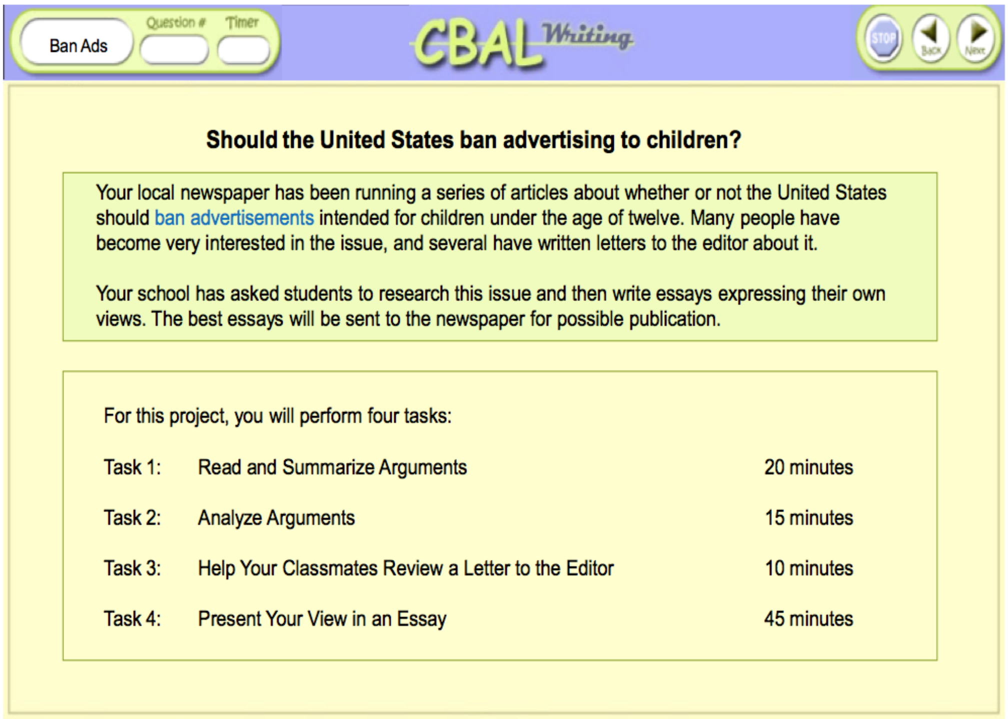
In addition to the final essay task, each CBAL writing assessment begins with a lead-in section in which students complete a range of selected-response and short writing tasks. Lead-in tasks are selected to measure critical reading, writing and thinking abilities related to the final writing task. For example, when the essay task focuses on building arguments about an issue, writers are asked to complete shorter tasks relevant to completing the essay, including:

- summarizing articles about the issue;
- completing selected-response questions probing their critical understanding of arguments on either side of the issue; and
- preparing a short critique of a straw-man argument on the same issue.

The lead-in tasks directly assess necessary component skills, while also modeling appropriate strategies for writing from sources, and thus help disambiguate whether students are able to think about the final writing prompt in task-appropriate ways. The selection of lead-in tasks varies by genre, and provides a way to target specific component skills that may be relatively difficult to assess from the final written product.

To give a sense of how these strategies are coordinated in CBAL assessment designs, Figure 1 gives a screenshot of the opening screen of one assessment, showing the overall design of the test - in particular, its organization into a scenario structure in which the lead-in tasks are intended to help prepare students for the final writing task. Figure 2 shows one of the lead-in tasks for the same test, and Figure 3 shows the instruction screen for the essay task.

.



**Should the United States ban advertising to children?**

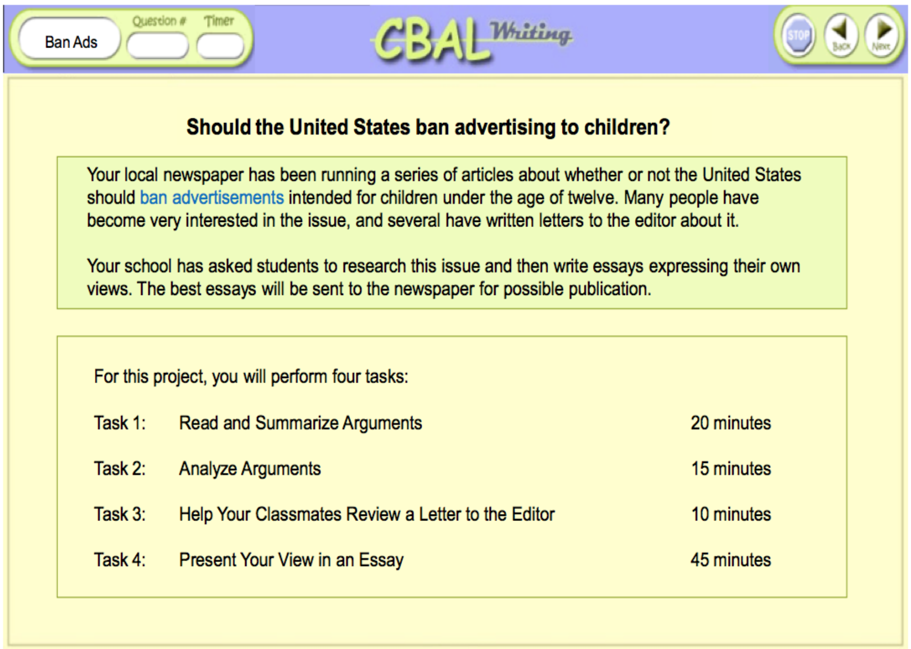
Your local newspaper has been running a series of articles about whether or not the United States should ban advertisements intended for children under the age of twelve. Many people have become very interested in the issue, and several have written letters to the editor about it.

Your school has asked students to research this issue and then write essays expressing their own views. The best essays will be sent to the newspaper for possible publication.

For this project, you will perform four tasks:

| | | |
|---|---|---|
| Task 1: | Read and Summarize Arguments | 20 minutes |
| Task 2: | Analyze Arguments | 15 minutes |
| Task 3: | Help Your Classmates Review a Letter to the Editor | 10 minutes |
| Task 4: | Present Your View in an Essay | 45 minutes |

Figure 2. Sample lead-in task for the "Ban Ads" Assessment



**Should the United States ban advertising to children?**

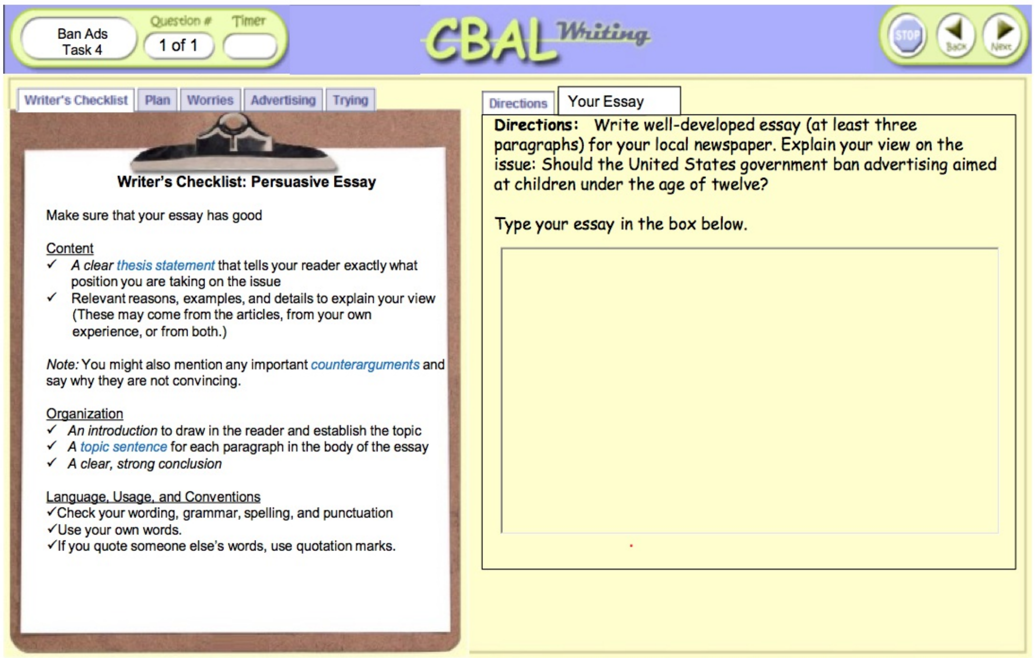
Your local newspaper has been running a series of articles about whether or not the United States should ban advertisements intended for children under the age of twelve. Many people have become very interested in the issue, and several have written letters to the editor about it.

Your school has asked students to research this issue and then write essays expressing their own views. The best essays will be sent to the newspaper for possible publication.

For this project, you will perform four tasks:

| | | |
|---|---|---|
| Task 1: | Read and Summarize Arguments | 20 minutes |
| Task 2: | Analyze Arguments | 15 minutes |
| Task 3: | Help Your Classmates Review a Letter to the Editor | 10 minutes |
| Task 4: | Present Your View in an Essay | 45 minutes |

Figure 3. Essay screen for the "Ban Ads" Assessment. The tabs allow students to access a planning tool and the three articles they read while they were completing the lead-in tasks.



**Writer's Checklist: Persuasive Essay**

Make sure that your essay has good

Content

- ✓ A clear *thesis statement* that tells your reader exactly what position you are taking on the issue
- ✓ Relevant reasons, examples, and details to explain your view (These may come from the articles, from your own experience, or from both.)

*Note:* You might also mention any important *counterarguments* and say why they are not convincing.

Organization

- ✓ *An introduction* to draw in the reader and establish the topic
- ✓ A *topic sentence* for each paragraph in the body of the essay
- ✓ *A clear, strong conclusion*

Language, Usage, and Conventions

- ✓Check your wording, grammar, spelling, and punctuation
- ✓Use your own words.
- ✓If you quote someone else's words, use quotation marks.

**Directions:** Write well-developed essay (at least three paragraphs) for your local newspaper. Explain your view on the issue: Should the United States government ban advertising aimed at children under the age of twelve?

Type your essay in the box below.

### *Measuring use of sources.*

When people write from sources, they must incorporate content from the source in their written response. If they do so carelessly, the result may be plagiarism, or at least a lack of any original material that goes beyond simple summarization. Problems in the use of source materials may reflect various factors, such as inadequate reading comprehension or difficulties in text production that make it hard to combine comprehension with productive activities such as paraphrase or summarization. In CBAL writing assessments, a limited number of sources are directly provided as part of the test. Some aspects of the way writers handle sources, such as the extent to which they include quoted material, and whether or not they mark them as quotations, can be directly (and automatically) measured. Existing AES systems do not generally include such features in the core engine, because the typical on-demand writing assessment does not involve writing from sources.

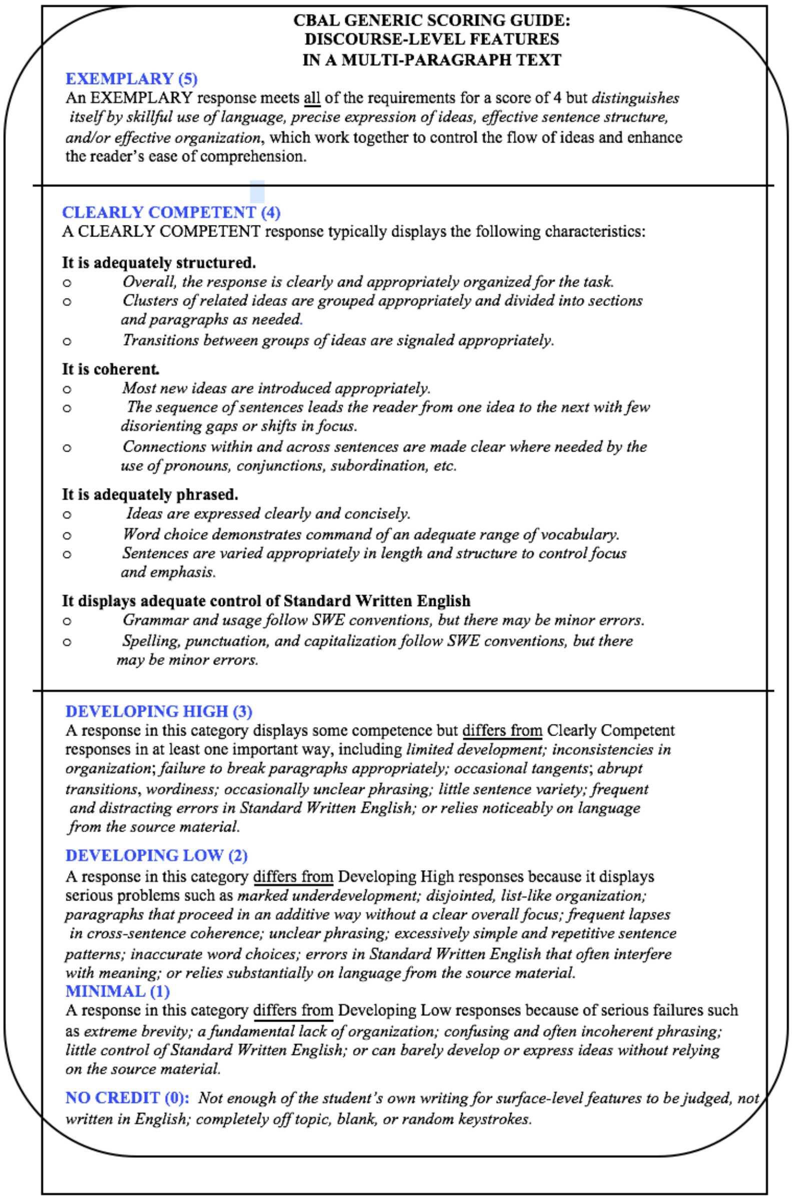
### *Keystroke logs.*

CBAL writing assessments are administered online, and so it is possible to collect a keystroke log capturing the exact time course of text production. We can extract information about a variety of event types from the keystroke log, such as the length of bursts of text production, the amount of time spent in editing events (such as cuts, pastes, jumps, and backspacing), and latency at various junctures (between characters, between words, between sentences). The literature suggests that such features provide evidence about aspects of the writing process such as fluency and control of text production, editing and revision behaviors, and time spent on planning (Miller, Lindgren, & Sullivan, 2008; Wengelin et al., 2009), which may help disambiguate the extent to which a writer's efforts are going into planning, revision, and text production. Prior work on use of keystroke logging features to support innovative forms of writing assessment has indicated that keystrokes can be successfully captured, and that when captured, they provide useful information that contributes to prediction of test score above and beyond other variables (Almond, Deane, Quinlan, & Wagner, in press; Deane & Quinlan, 2010; Deane, Quinlan, & Kostin, 2011)

***Trait scoring to create a division of labor between human and automated scoring.***

Current-generation AES engines provide direct evidence only with respect to some aspects of writing skill, typically involving features that measure such things as basic essay structure, vocabulary, style, grammar, usage, mechanics, and spelling (Shermis, Burstein & Leacock, 2006; Shermis & Hammer, 2012). Humans are sensitive to such features, but are also able to focus on other, critical aspects of writing skill, such as quality of argumentation or effectiveness of textual analysis. Logically, automated scoring could be substituted for human scoring of equivalent traits - namely, those traits for which automated scoring provides the best measurement - while making use of human scoring for those traits that are less appropriately scored by a machine. CBAL writing assessments are therefore scored using trait-based rubrics for each essay task. These rubrics distinguish between content- and genre- based elements that require social and conceptual reasoning (such as quality of argumentation) and more generic skills that are susceptible to being scored with an automated scoring engine (such as evaluation of grammar and spelling). Figures 4 and 5 illustrate two of the rubrics that have been developed for CBAL writing assessments: a common writing quality rubric (Figure 4) and a rubric focused on strength of argumentation (Figure 5). For comparison, genre-specific rubrics for literary interpretation (as in Mango Street) and arguing using criteria (as in Service Learning) are presented in Figures 6 and 7. Our expectation is that an automated scoring engine can be trained to predict the kinds of judgments required in the rubric shown in Figure 4, allowing substitution of automated scoring for human scoring for these traits. The features used by the e-rater scoring engine, in particular, is closely aligned to the elements included in the common writing quality rubric, which include features intended to measure organization, development, mechanics, vocabulary, grammar, usage, and style. This leaves human scorers to focus on the relatively difficult aspects of writing that are critical to successful performance, but which cannot be machine-scored because they focus on deeper forms of reasoning (as in the rubric shown in Figures 5-7).

Figure 4. Generic rubric focusing on print, verbal and discourse features.



**CBAL GENERIC SCORING GUIDE:**
**DISCOURSE-LEVEL FEATURES**
**IN A MULTI-PARAGRAPH TEXT**

**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.*

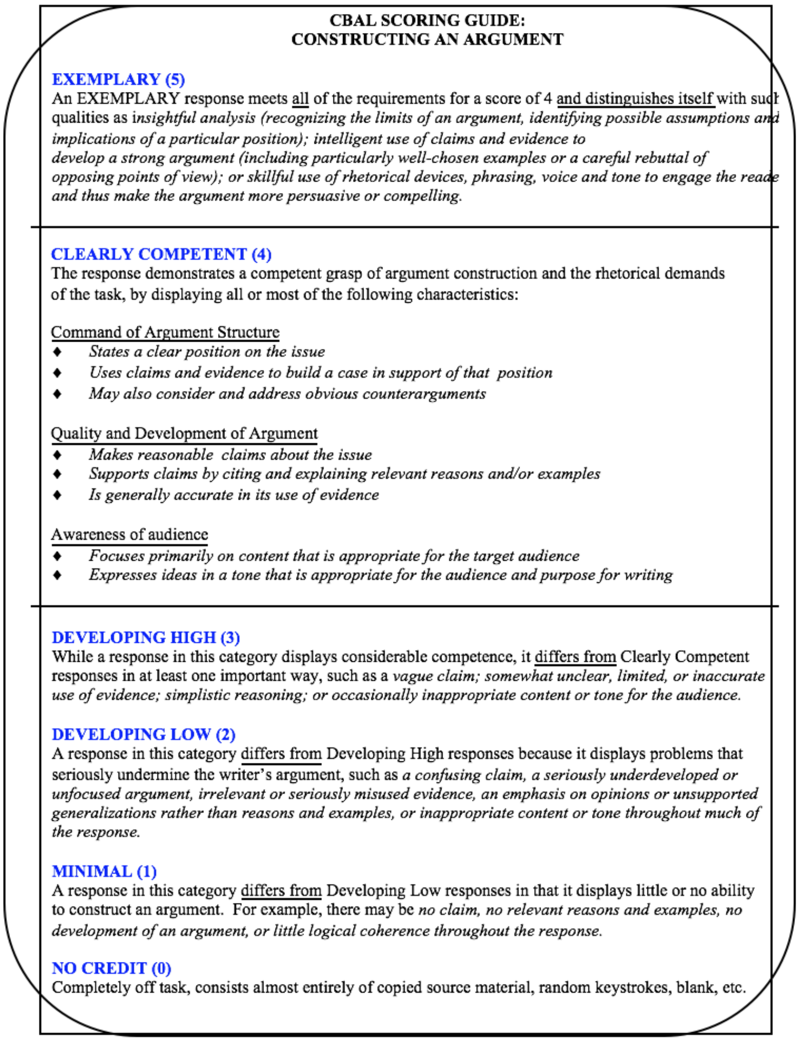
**DEVELOPING HIGH (3)**
A response in this category displays some competence but differs from Clearly Competent responses in at least one important way, including *limited development; inconsistencies in organization*; *failure to break paragraphs appropriately; occasional tangents*; *abrupt transitions*, *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 5. Rubric for "Ban Ads" test form focusing on rhetorical effectiveness and quality of argumentation.



**CBAL SCORING GUIDE:**
**CONSTRUCTING AN ARGUMENT**

**EXEMPLARY (5)**
An EXEMPLARY response meets all of the requirements for a score of 4 and distinguishes itself with such qualities as i*nsightful 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 reade 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*

Quality and Development of Argument

- *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*

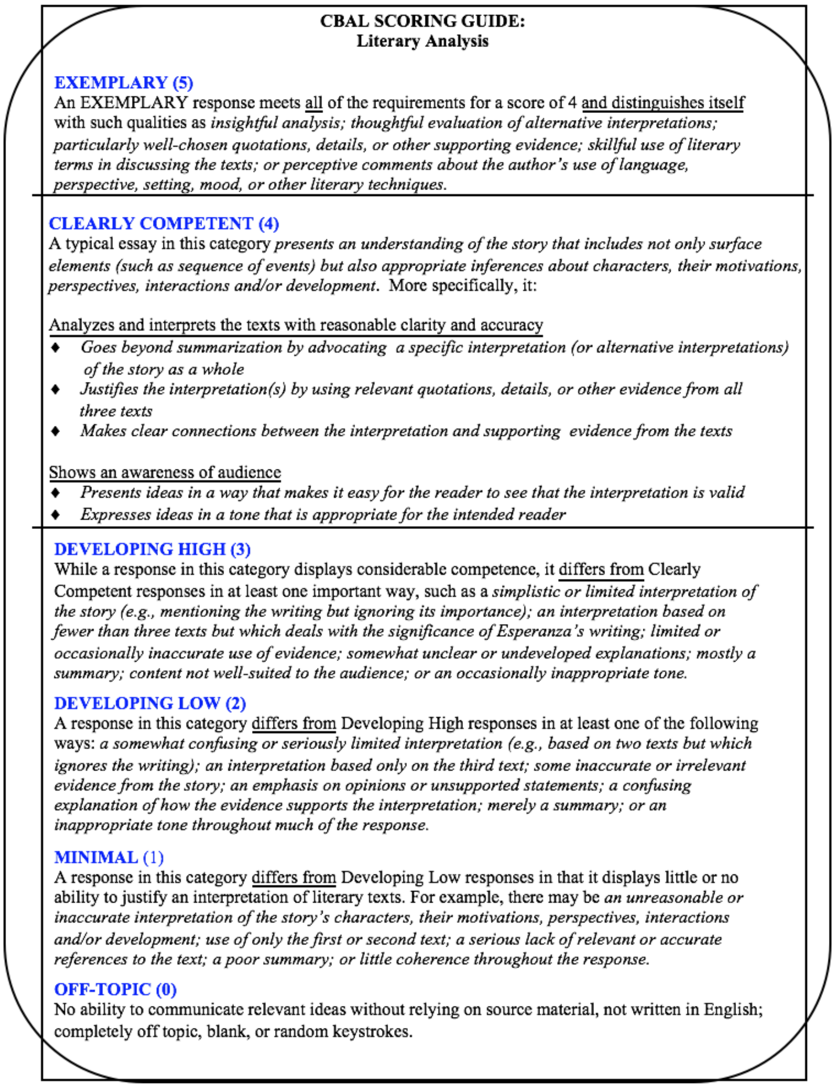
**DEVELOPING HIGH (3)**
While a response in this category displays considerable competence, it differs from Clearly Competent responses in at least one important way, such as a *vague claim; somewhat unclear, limited, or inaccurate use of evidence; simplistic reasoning; or occasionally inappropriate 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 6. Genre-Specific Rubric for Literary Analysis (Applied to essays written to the Mango Street prompt)



**CBAL SCORING GUIDE:**
**Literary Analysis**

**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*

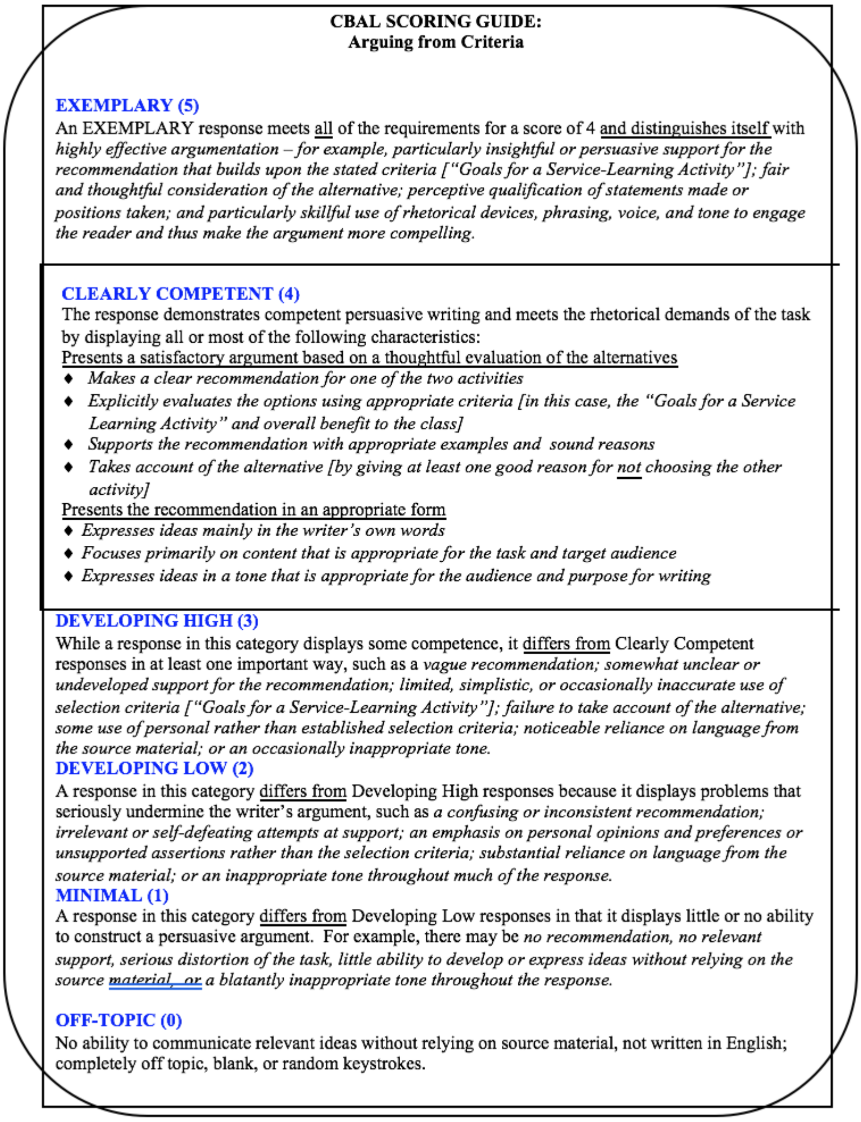
**DEVELOPING HIGH (3)**
While a response in this category displays considerable competence, it differs from Clearly Competent responses in at least one important way, such as a *simplistic or limited interpretation of the story (e.g., mentioning the writing but ignoring its importance); an interpretation 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 7. Genre-Specific Rubric for Arguing from Criteria (Applied to essays written to the Service Learning prompt)



**CBAL SCORING GUIDE:**
**Arguing from Criteria**

**EXEMPLARY (5)**
An EXEMPLARY response meets all of the requirements for a score of 4 and distinguishes itself with *highly effective argumentation – for example, particularly insightful or persuasive support for the recommendation that builds upon the stated criteria ["Goals for a Service-Learning Activity"]; fair and thoughtful consideration of the alternative; perceptive qualification of statements made or positions taken; and particularly skillful use of rhetorical devices, phrasing, voice, and tone to engage the reader and thus make the argument more compelling.*

**CLEARLY COMPETENT (4)**
The response demonstrates competent persuasive writing and meets the rhetorical demands of the task by displaying all or most of the following characteristics:
Presents a satisfactory argument based on a thoughtful evaluation of the alternatives
- ♦ *Makes a clear recommendation for one of the two activities*
- ♦ *Explicitly evaluates the options using appropriate criteria [in this case, the "Goals for a Service Learning Activity" and overall benefit to the class]*
- ♦ *Supports the recommendation with appropriate examples and sound reasons*
- ♦ *Takes account of the alternative [by giving at least one good reason for not choosing the other activity]*

Presents the recommendation in an appropriate form
- ♦ *Expresses ideas mainly in the writer's own words*
- ♦ *Focuses primarily on content that is appropriate for the task and 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 category displays some competence, it differs from Clearly Competent responses in at least one important way, such as a *vague recommendation; somewhat unclear or undeveloped support for the recommendation; limited, simplistic, or occasionally inaccurate use of selection criteria ["Goals for a Service-Learning Activity"]; failure to take account of the alternative; some use of personal rather than established selection criteria; noticeable reliance on language from the source material; or an occasionally inappropriate tone.*

**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 or inconsistent recommendation; irrelevant or self-defeating attempts at support; an emphasis on personal opinions and preferences or unsupported assertions rather than the selection criteria; substantial reliance on language from the source material; 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 construct a persuasive argument. For example, there may be *no recommendation, no relevant support, serious distortion of the task, little ability to develop or express ideas without relying on the source material, or a blatantly inappropriate tone 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.

***The place of automated essay scoring.***
As this discussion suggests, CBAL writing assessments are designed to provide multiple converging strands of evidence rich enough to address many different aspects of writing skill. AES technologies fit into this ecology by providing a critical source of evidence about student performance on a range of traits susceptible to direct measurement from the final written response. The most direct implementation of automated scoring in this context would exploit existing statistical training methods by training an AES model against human scores - in this case, scores focused on the specific traits addressed by the scoring engine. This is the application that we explore in this paper.

In particular, we examine the following questions:

1. *Accuracy of the scoring model.* Can we build accurate automated essay scoring models for writing prompts in an innovative assessment of writing from sources?
2. *Validation.* When we can build such models, how valid are the results? What limits exist on their use and interpretation?
3. *Usefulness.* More generally, how much can AES technology contribute to an assessment design that draws upon multiple sources of evidence to assess writing proficiency?

## Method

### Participants and Administration Procedures

In fall 2009, CBAL English Language Arts assessments were administered to a convenience sample of 2,606 8th-grade students attending 24 schools spread across 18 U.S. states. Each student took two of four writing test forms. In 2011, CBAL English Language Arts assessments were administered to a second convenience sample of 3,667 8th-grade students from 35 schools across the U.S. Each student was assigned two forms randomly selected from a set containing two reading assessments and two writing assessments; 2,247 students completed at least one writing form, and 596 completed both. The reading and writing tests included forms that focused on informational texts (the 'Wind Power' form for reading, and the 'Ban Ads' form for writing, so named for the topic that was the focus of the test scenario) and on literary texts (The 'Seasons' form for reading, and the 'Mango Street' form for writing).

In both the 2009 and 2011 administrations, the schools represented a mix of urban, suburban, and rural institutions from across the U.S. Each student took two of the four forms within a one month period (in the 2009 study) or a two week period (in the 2011 study). Each possible combination and order of forms was used and the forms were randomly assigned to students within schools, but constrained so that the same form was never administered twice in the same school.

**Instruments**

In the 2009 administration, four writing assessments were administered (Deane, 2010; Deane et al., 2010; Fu, Chung, & Wise, 2013; Fu & Wise, 2012). In spring 2011, two of these assessments were administered a second time (Cline, 2012). Each test form contained a set of lead-in tasks, which students were allowed 45 minutes to complete, and an essay-writing task, also 45 minutes.

In the 2009 study, two ETS raters scored each essay using the common rubric (see Figure 4), and two other ETS raters scored it using a genre-specific rubric focused on critical thinking skills such as argumentation (see Figure 5) or literary analysis. In the 2011 study, all essays were scored by at least two raters (one per rubric), but 20% of the essays were double-scored on both rubrics. The remaining 80% of essays were scored by two raters, one rater per rubric. Several raters were involved in each scoring effort. The four forms focused on the following genres.

***Argumentation.***

Argumentation addressed the issue of whether advertisements to children under age 12 should be banned. Below, this assessment is identified as "Ban Ads." It contained one essay item, three short constructed-response items, and 21 selected-response items. Students read three articles on the issue, summarized the texts, analyzed the arguments, and wrote an essay.

***Literary Analysis.***

For the literary analysis assessment, students were required to analyze three excerpts from the novel, The House on Mango Street. Below, this assessment is identified as "Mango Street." It contained one essay item, two short constructed-response items, and ten selected-response items. Students read the excerpts, analyzed and responded to interpretations of them, and wrote an essay.

***Policy Recommendation.***

Students were required to recommend one classroom service-learning project over another, using pre-established criteria for evaluation. Below, this assessment is identified as "Service Learning." It contained one essay item, one short constructed-response item, and 15 selected-response items. Students read descriptions of projects and a document indicating criteria that should be met by a good service-learning project. They then analyzed projects using the criteria, considered how to improve a project, and wrote up a recommendation.

***Informational Pamphlet.***

This assessment required students to write one section of an informational pamphlet based upon notes from research about invasive plant species. It contained one long and one short constructed-response writing item and 30 selected-response items. Students read articles on the subject, analyzed research notes and organized them for a pamphlet, after which they wrote one section in a partially completed pamphlet. Initial analysis reported in Deane (2010) indicated that the long writing task in this form (nicknamed "Invasive Species" after the topic to which it was applied) was problematic, in part due to widespread inappropriate use of cut-and-paste for research information in the essay task, and it was excluded from further study.

***Dataset.***

Table 1 provides an overview of the numbers of students that took each form. The tests were given in two sessions, so there was some attrition between the session in which students took the lead-in task and the session in which they wrote the essay. When automated scoring models were prepared and tested, essays that were blank, not in English, or too short for machine scoring according to operational standards for e-rater (i.e., less than 50 words) were removed from the set. There were somewhat more problematic responses of this type in the 2009 dataset. Since such responses are identified by pre-filters, and excluded from automated scoring in operational use in the automated essay scoring engine used in this study, they need to be excluded from the engine training and cross-validation sets, as well.

Table 1.

*Number of Students Attempting Each Test Form*

| | | N Attempting Lead-In tasks | N Attempting Essay | N essays used for e-rater analysis |
|---|---|---|---|---|
| Ban Ads | 2009 Study | 1,160 | 1,054 | 875 |
| | 2011 Study | 1,718 | 1,446 | 1,403 |
| Mango Street | 2009 Study | 1,213 | 1,109 | 940 |
| | 2011 Study | 1,688 | 1,616 | 1,539 |
| Service Learning | 2009 Study | 1,193 | 1,116 | 950 |

## Data Analysis

### *Accuracy of the automated scoring model.*

In the 2009 study, all four forms were analyzed to estimate score reliability, determine inter-rater agreement, and appraise the accuracy of AES models - specifically, to evaluate models built using ETS's e-rater® essay scoring engine (Attali & Burstein, 2009). The same analyses were repeated for the two forms used in the 2011 study.

We hypothesized that e-rater was best suited to predicting human scores for the common rubric that focused on such features as organization, development, word choice, and adherence to conventions. e-rater models were therefore built to predict scores on this rubric. As a result, scores on the other (genre- and prompt-specific) rubrics were used in later stages of the analysis for purposes of validation.

We calculated the following statistics to evaluate levels of inter-rater agreement for the human raters: 1. Means and standard deviations for the first and second rater; 2. Measures of agreement, including correlations, standardized differences, kappa, weighted kappa, exact agreement, and adjacent agreement. These statistics were used to establish a baseline level of human performance. Established operational standards at ETS were used to evaluate the levels of inter-rater agreement displayed in each study (Williamson, Xi, & Breyer, 2012).

e-rater models were built using the following procedure:

1. Essay responses from the 2009 data set were randomly divided into equal model-building (MB) and cross-validation (XV) subsets, separately by prompt for Ban Ads, Mango Street, and Service Learning.
2. Model-building sets were used to train an AES (e-rater) model for each prompt, and cross-validation subsets were used to evaluate the model that resulted.

The evaluation of e-rater models was conducted by comparing actual and predicted human scores and by estimating the agreement between e-rater and the first human rater. The following statistics were calculated:

1. For both datasets, means and standard deviations for human raters and for e-rater predicted scores.
2. For the 2009 dataset, average human scores - i.e., the mean score for each essay, averaged across both raters. These average human scores were used to train the e-rater model. Means and standard deviations were calculated both for individual raters and for the average scores.
3. For the 2009 dataset, statistics measuring agreement between average human scores and e-rater, including
   a. the correlation between average human and predicted scores,
   b. the difference between the average human and predicted means, and
   c. mean squared error.
4. For both datasets, statistics measuring agreement between the first human rater, the second human rater, and e-rater. The following statistics were calculated:
   a. correlations,
   b. the kappa and weighted kappa coefficients, and
   c. adjacent and exact agreement.

Established operational standards at ETS were used to evaluate how accurately e-rater predicted human scores.

### *Validity evidence.*

In accordance with the standards for validity specified by the Standards for Psychological and Educational Testing (AERA, APA, &

NCME, 1999) we would expect that e-rater models, to the extent that they captured the same construct as the common writing-quality rubric, would show similar relationships to other measures. Such predictive relations are not the only source of evidence, but are nonetheless critical in building a validity argument. In these studies, we have several comparisons available.

*Comparisons across the three prompts in the 2009 study.*
Two of the prompts (Ban Ads and Service Learning) focused on argument-based informational writing. The third (Mango Street) focused on literary analysis. We would expect that scores generated from the Ban Ads and Service Learning e-rater models would reflect models whose feature weights were more similar to one another than either is to the feature weights assigned by the Mango Street e-rater model.

*Correlations with human scores on the genre-specific rubrics (argumentation for Ban Ads; literary interpretation for Mango Street).*
We would expect moderate to strong correlations between human scores on the common rubric and human scores on the genre-specific rubrics, since they measure different but related aspects of the same performance, and we would expect e-rater predicted scores to show similar patterns of correlation.

*Correlations with total scores on the lead-in tasks in each test.*
We would expect moderate correlations between the lead-in tasks and predicted essay scores, since the lead-in tasks are designed to measure skills not perfectly measured in the final written product.

*Correlations with total scores on the reading tests in the 2011 study.*
We would expect moderate correlations between human writing and reading scores, since these are related but distinct skills, and we would expect a similar pattern of correlations between e-rater predicted scores and reading scores.

*Correlations with keystroke log features, such as the latency and variance of pauses within and between words.*
We would expect weak to moderate correlations with individual features, since those features measure aspects of a very different aspect of writing than are addressed by human scores for the final written product. Once again, we expect that e-rater scores will show comparable correlation patterns.

*Usefulness.*
To be useful, automatically predicted scores need not account for every aspect of the writing construct, but they must account for a large, significant portion of total test score to be meaningfully combined with other items on the same test. The usefulness of AES will be supported if it can be used reliably as the primary predictor, using other measures to identify cases where students depart from the main trend line. This question is addressed by examining correlations with total test score and by constructing regression models that predict total test score using AES features as a predictor

## Results

**Form Reliability and Accuracy of Human Scoring**
Despite the difficulties observed with the Invasive Species form, all four tests performed well psychometrically. Score reliabilities (Cronbach's alpha coefficients for the entire test form), for example, ranged from .76 to .86 (Bennett, 2011).

Tables 2 and 3 show how consistently the human raters performed in the 2009 and 2011 administration when assessed using the standard methods employed for e-rater evaluation (and excluding from consideration those essays that would be flagged and therefore removed from the dataset prior to e-rater model training).

Table 2

*Comparison of Human Raters on the 2009 Cross-validation Set (standard differences greater than the threshold for operational use are marked in gray.)*

| | | Human1 | | Human2 | | Statistics | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Prompt | N | Mean | Std. Dev. | Mean | Std. Dev. | Std. Diff. | Kappa | Wtd Kappa | % Exact Agreement | % Adj. Agreement | Corre-lation |
| Ban Ads | 448 | 2.39 | 0.99 | 2.19 | 1.08 | -0.20 | 0.43 | 0.73 | 58.0 | 94.4 | 0.75 |
| Mango Street | 489 | 2.24 | 1.03 | 2.36 | 1.06 | 0.12 | 0.38 | 0.72 | 54.4 | 94.9 | 0.72 |
| Service Learning | 488 | 2.53 | 1.07 | 2.71 | 1.00 | 0.17 | 0.34 | 0.72 | 51.4 | 96.1 | 0.73 |
| Mean | 475.0 | 2.39 | 1.03 | 2.42 | 1.05 | 0.03 | 0.38 | 0.72 | 54.6 | 95.1 | 0.73 |

Table 3

*Comparison of Human Raters on the 2011 Dataset*

| | | Human1 | | Human2 | | Statistics | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Prompt | N | Mean | Std. Dev. | Mean | Std. Dev. | Std. Diff. | Kappa | Wtd Kappa | % Exact Agreement | % Adj. Agreement | Corre-lation |
| Ban Ads | 283 | 2.30 | 1.04 | 2.28 | 0.98 | -0.03 | 0.60 | 0.83 | 70.7 | 98.2 | 0.83 |
| Mango Street | 301 | 2.17 | 0.96 | 2.15 | 0.91 | -0.03 | 0.70 | 0.86 | 79.4 | 98.7 | 0.86 |
| Mean | 292 | 2.24 | 1.00 | 2.21 | 0.95 | -0.03 | 0.65 | 0.84 | 75.0 | 98.5 | 0.85 |

These are results prior to adjudication. Operational standards for an automated scoring model at ETS require that the standardized differences among human raters be less than .15, and that the weighted kappa and correlation between the two scorers both be above .70. When the raters for the 2009 study are evaluated against the cross-validation set, the standardized differences are slightly higher than ideal for operational use (between .15 and .20), though the weighted kappa and correlations are above the required thresholds. The performance of raters in the 2011 study fully meets the requirements for operational use. There are no standardized differences between the two rater distributions greater than 3%; weighted kappas are all over .80, as are the correlations, and exact agreement is greater than 70%.

The general pattern of performance thus suggests improvements in the quality of human scoring from 2009 to 2011, although double-scored sets are relatively small. This finding is consistent with efforts that were made to improve the instruction and training given to the 2011 raters in the light of experience with the 2009 scoring effort.

In the present study, the 2009 scores were used to train and cross-validate an e-rater model, which was then applied to the 2011 data. This use risks some decrements to the quality of the e-rater model, but provides a more rigorous test of how well the system performs when generalized to the 2011 dataset.

### Accuracy of the Automated Scoring Model

***Type of model built.***

E-rater supports two kinds of models: so-called prompt-specific models, in which content vector analysis (CVA) is applied to identify whether an essay has vocabulary characteristic of high-scoring essays written to the same prompt, and generic models, which do not use prompt-specific features, and can therefore be applied to essays written to a variety of different topics. With data from only three prompts available, it was not possible to construct true generic models. It was, however, possible to evaluate whether the CVA features made a significant difference in model performance. Models with and without CBA features were built, but the differences between them were very small (degradation in weighted kappa of less than .02 when generic models are substituted for prompt-specific models). In the discussion that follows, therefore, we report only the results for e-rater models without CVA features. We hope in future research to determine whether these models can be applied across a range of different topics, yielding true generic models.

***Agreement of human raters and e-rater in the 2009 dataset.***

On the 2009 cross-validation set, differences between the average human score and the predicted e-rater score were generally small (standard difference < -.006 for Mango Street and < -.03 for Ban Ads) though rather larger for Service Learning (standard difference = -.09). Mean squared error fell in the range .29 to .38, and the correlations between human and predicted e-rater scores were uniformly above 0.80 (see Table 4). These results all fall within the acceptable range, indicating reasonable fit between human scores and e-rater predictors.

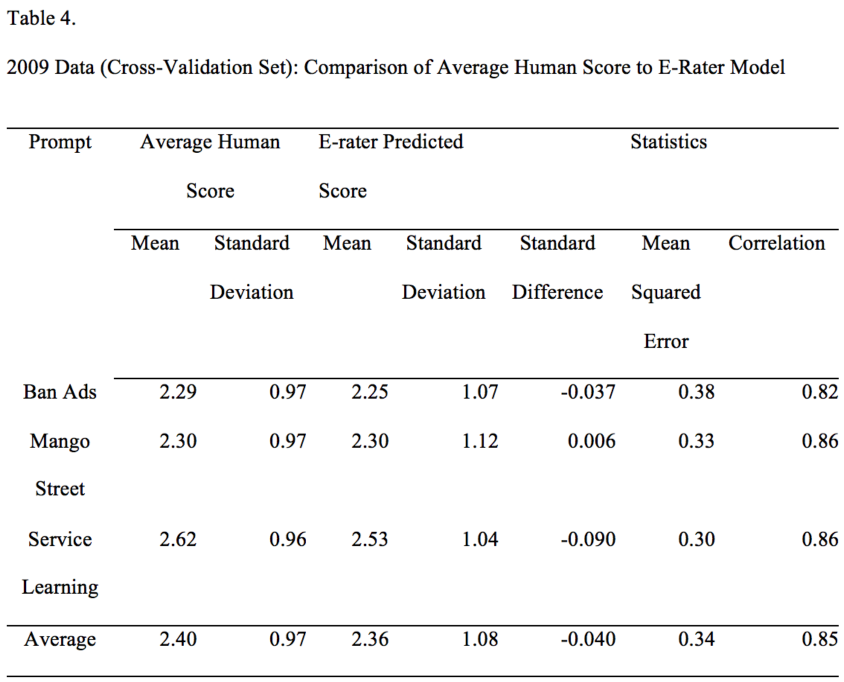


Table 4.

2009 Data (Cross-Validation Set): Comparison of Average Human Score to E-Rater Model

| Prompt | Average Human Score | | E-rater Predicted Score | | Statistics | | |
|---|---|---|---|---|---|---|---|
| | Mean | Standard Deviation | Mean | Standard Deviation | Standard Difference | Mean Squared Error | Correlation |
| Ban Ads | 2.29 | 0.97 | 2.25 | 1.07 | -0.037 | 0.38 | 0.82 |
| Mango Street | 2.30 | 0.97 | 2.30 | 1.12 | 0.006 | 0.33 | 0.86 |
| Service Learning | 2.62 | 0.96 | 2.53 | 1.04 | -0.090 | 0.30 | 0.86 |
| Average | 2.40 | 0.97 | 2.36 | 1.08 | -0.040 | 0.34 | 0.85 |

[4] H2/e-rater correlations are very similar to H1/e-rater correlations, and so are not shown here.

In the cross-validation data from the 2009 study, the e-rater model agrees more closely with the first human rater than two human raters agree with one another. Tables 5 through 7 illustrate the pattern, outlining the performance of humans and e-rater on each prompt in the 2009 cross-validation set. The differences are not particularly great, though in some cases (e.g., Service Learning) they are large enough to suggest that improvements in training of human scorers might be appropriate. However, the agreement between human raters and e-rater remains above the threshold for operational deployment (e.g., weighted kappa > 0.70).

Table 5

*Ban Ads Prompt (2009 Dataset): Agreement Rates Between the First Human Rater, Second Human Rater, and E-Rater (Cross-validation Set)*

| Measure | H1 / H2 | H1 /E-rater[4] | Difference |
|---|---|---|---|
| Kappa | 0.43 | 0.44 | -0.01 |
| Wtd Kappa | 0.73 | 0.76 | -0.03 |
| % exact Agreement | 58.0 | 59.4 | -1.3 |
| % adjacent Agreement | 94.4 | 97.8 | -3.4 |
| Correlation | 0.75 | 0.76 | -0.01 |

Table 6

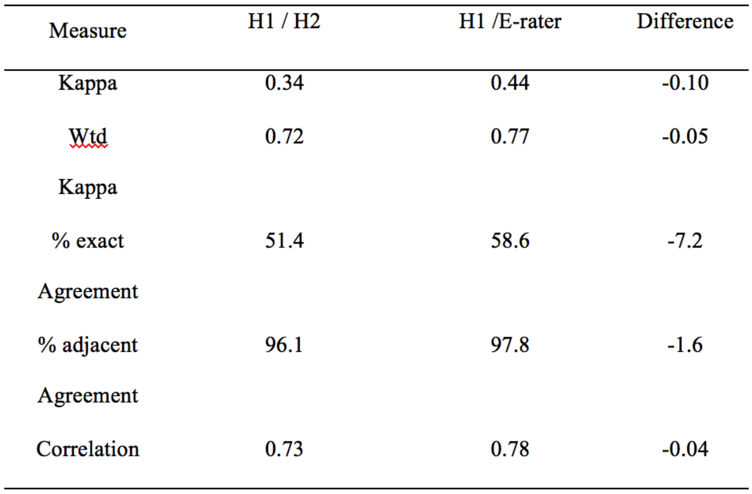
*Mango Street Prompt (2009 Dataset): Agreement Rates Between the First Human Rater, Second Human Rater, and E-Rater (Cross-validation Set).*

| Measure | H1 / H2 | H1 /E-rater[5] | Difference |
|---|---|---|---|
| Kappa | 0.38 | 0.45 | -0.07 |
| Wtd Kappa | 0.72 | 0.77 | -0.05 |
| % exact Agreement | 54.4 | 59.5 | -5.1 |
| % adjacent Agreement | 94.9 | 97.3 | -2.5 |
| Correlation | 0.72 | 0.78 | -0.06 |

[5] H2/e-rater correlations are very similar to H1/e-rater correlations, and so are not shown here.

Table 7

*Service Learning (2009 dataset): Agreement rates between the first human rater, second human rater, and e-rater model (cross-validation set)*



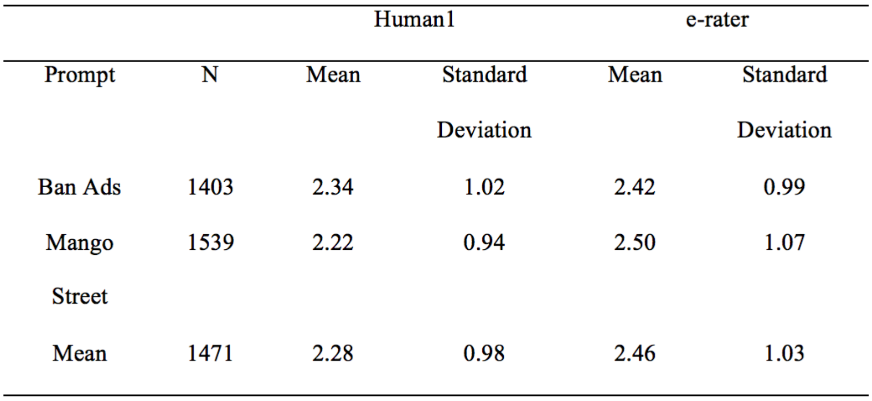
| Measure | H1 / H2 | H1 /E-rater | Difference |
|---|---|---|---|
| Kappa | 0.34 | 0.44 | -0.10 |
| Wtd Kappa | 0.72 | 0.77 | -0.05 |
| % exact Agreement | 51.4 | 58.6 | -7.2 |
| % adjacent Agreement | 96.1 | 97.8 | -1.6 |
| Correlation | 0.73 | 0.78 | -0.04 |

***Agreement of human raters and e-rater in the 2011 dataset.***

When we apply the model built on 2009 data to the 2011 dataset, we obtain very different results for the Mango Street and Ban Ads prompts. The general pattern of performance on the Ban Ads prompt is consistent with the 2009 model, whereas there is a significant difference in the pattern of performance on the Mango Street prompt. Tables 8 and 9 show how the e-rater model compares with human ratings when applied to the entire 2011 dataset. Even though all the other statistics are in the normal range, and above the thresholds for operational performance (e.g., weighted kappa and human/e-rater correlation > .70), the standardized difference between human ratings and the e-rater scores for Mango Street is .28, well above ETS's .15 threshold for acceptable model performance. This result indicates that there are differences either in the distribution of responses or in raters' scoring practices between the 2009 and 2011 datasets.

Table 8

*2011 data: means and standard deviations for the first human rater and the e-rater model*



| | | Human1 | | e-rater | |
|---|---|---|---|---|---|
| Prompt | N | Mean | Standard Deviation | Mean | Standard Deviation |
| Ban Ads | 1403 | 2.34 | 1.02 | 2.42 | 0.99 |
| Mango Street | 1539 | 2.22 | 0.94 | 2.50 | 1.07 |
| Mean | 1471 | 2.28 | 0.98 | 2.46 | 1.03 |

Table 9

*2011 Data: Agreement Between the First Human Rater and the E-Rater Model. Standard differences greater than the operational threshold are highlighted in gray.*

| Prompt | Std Difference | Kappa | Wtd Kappa | % Exact Agreement | % Adjacent Agreement | Correlation |
|---|---|---|---|---|---|---|
| Ban Ads | 0.08 | 0.57 | 0.83 | 68.2 | 99.2 | 0.83 |
| Mango Street | 0.28 | 0.47 | 0.75 | 60.0 | 96.5 | 0.79 |
| Average | 0.18 | 0.52 | 0.79 | 64.1 | 97.9 | 0.81 |

Williams and Breyer (2011) have examined this difference in detail, using e-rater as an anchor to equate the two sets. Their analysis suggests that when the two administrations are placed on a common scale, the students in the 2011 administration are on average more proficient than the students in the 2009 administration, but that the 2011 raters were more severe on both prompts. They report that this difference is much more striking with Mango Street than Ban Ads, even though both administrations have comparable score distributions.

However, rater training was revised between 2009 and 2011 precisely to address problems identified in the 2009 scoring, including a tendency of raters not to notice plagiarism from the stimulus materials. Thus, the relatively lenient scoring in 2009 might be due to rater failure to address all features of the rubric. Similarly, an increase in scoring quality might explain the increase in scoring consistency among raters for the 2011 study. To evaluate whether this hypothesis might be true, we had a group of human experts examine cases where there were discrepancies of more than 1.5 points between human and e-rater scores. Their evaluation indicated that that most of these discrepancies involved plagiarism from source documents, suggesting that the raters were more sensitive to this feature in 2011 than in 2009.

In a test of writing from sources, it is important to measure the extent to which students reproduce source materials inappropriately. Standard e-rater models do not directly incorporate measurement of plagiarism, though plagiarism detection software is used operationally to identify essays plagiarized from previous administrations or from the Internet. However, we were able to calculate a nonstandard feature that measures the percentage of a text copied from stimulus materials. As a result, we were able to assess whether plagiarism accounted for the discrepancy between humans and e-rater on the 2011 Mango Street administration. This feature identified the proportion of text copied verbatim, without surrounding quotation marks.

In the 2011 data, if we train an e-rater model to predict human generic-rubric scores, this feature is a significant predictor, with a beta weight of .09 and an $R^2$ change of +.01 for Ban Ads and a beta weight of .24 and an $R^2$ change of +.06 for Mango Street. The much larger impact of this feature upon the accuracy of an e-rater model for Mango Street suggests that the large discrepancy observed for Mango Street in Table 9 is probably due to an increased tendency for students to copy source materials verbatim. The Mango Street task in particular requires some use of quotations or paraphrases from the source materials to justify interpretations of the text, which may account for the greater weight attached to this feature in the Mango Street model.

To sum up: The performance of e-rater on the 2009 and 2011 datasets, for the most part, is at operational levels, though optimal performance may require enhancements to the model to account for plagiarism and other use of materials from sources. In principle, the difference between the Mango Street and Ban Ads models could be due to differences in genre (the type of writing required), or to differences in the choice of topic about which students are required to write. However, in the present study, there is no way to disambiguate differences between models, since the forms differ both in topic and in genre.

**Validity Evidence**

***Feature weights across e-rater models.***

In 2009 (and also in 2011), test forms were assigned in a counterbalanced fashion by school, subject to the constraint that no single test form could be administered in both testing windows at the same school. Within school and test sessions, the groups that took each test form were assigned randomly. Thus, it is reasonable to assume that each form was completed by equivalent subsamples (for more details, see Fu et al., 2012, 2013). We can therefore compare the feature weights in each e-rater model, with an eye toward determining whether the models are essentially similar. If all three tests evoke essentially similar kinds of writing performance, we would expect similar patterns of feature weights across models.

Table 10 captures the patterns of feature weights identified for each prompt. If we examine the relative magnitude of weights assigned to each feature, we see that they fall in exactly the same order for Ban Ads and Mango Street (Organization, Development, Mechanics, Average Word Length, Grammar, Style, Usage, Sophistication of Word Choice, Positive Feature), and that the third prompt, Service Learning, has the same order, but drops two features addressing idiomatic language and average word frequency.

Table 10: *Relative Weights of E-Rater Features across Models*

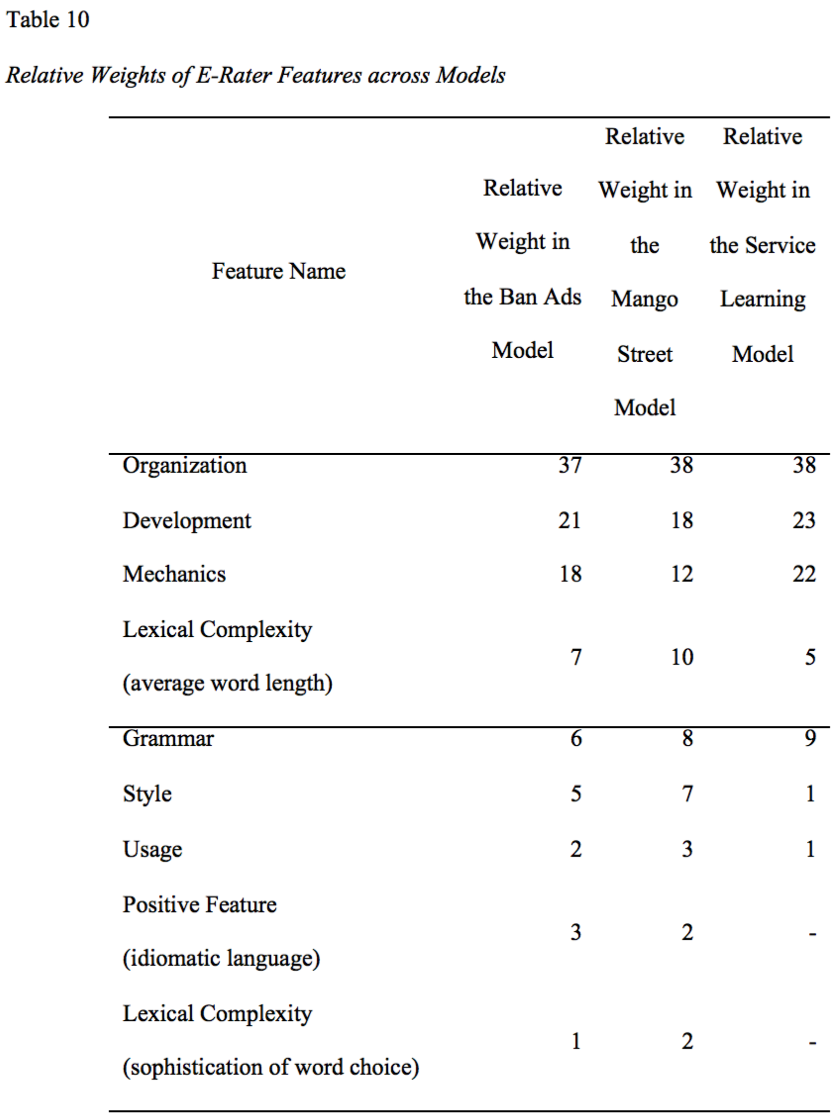


Table 10

*Relative Weights of E-Rater Features across Models*

| Feature Name | Relative Weight in the Ban Ads Model | Relative Weight in the Mango Street Model | Relative Weight in the Service Learning Model |
|---|---|---|---|
| Organization | 37 | 38 | 38 |
| Development | 21 | 18 | 23 |
| Mechanics | 18 | 12 | 22 |
| Lexical Complexity (average word length) | 7 | 10 | 5 |
| Grammar | 6 | 8 | 9 |
| Style | 5 | 7 | 1 |
| Usage | 2 | 3 | 1 |
| Positive Feature (idiomatic language) | 3 | 2 | - |
| Lexical Complexity (sophistication of word choice) | 1 | 2 | - |

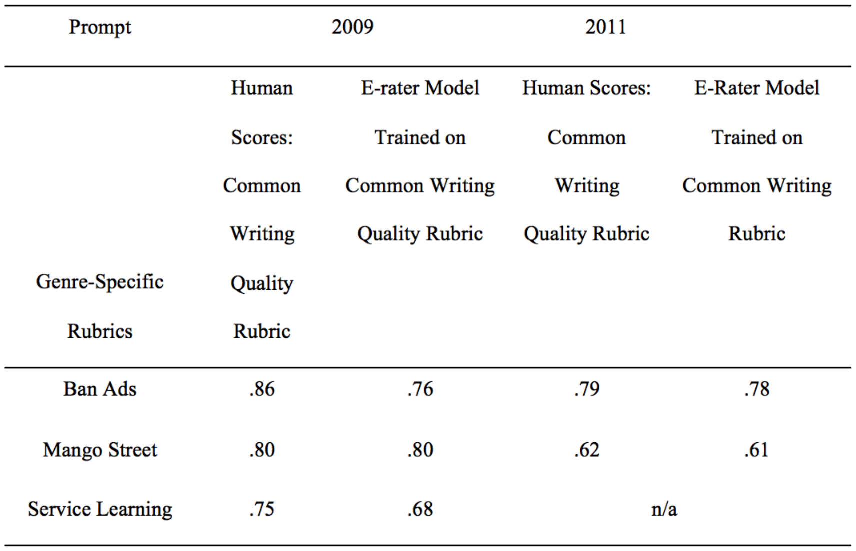
The underlying similarity of the models can be confirmed by examining the correlations among the standardized coefficients. When we do this, we obtain a pattern in which the Ban Ads and Service Learning models are very close (Pearson correlation = .99), but in which the Mango Street model is not quite so strongly correlated (Pearson correlation to Ban Ads = .88, and .87 to Service Learning). These results are consistent with the genre characteristics of each prompt: Both Ban Ads and Service Learning focus on persuasive writing, whereas the Mango Street prompt requires the writer to focus on literary interpretation and analysis.

***Comparison with human scores based on a genre-specific rubric.***

Unlike standard applications of AES, we have two sets of human scores available: One set that is based on the generic writing rubric (used to train the e-rater model), and another set based on a genre-specific rubric focused on quality of reasoning. The genre-specific rubrics focus on building arguments (for Ban Ads), applying evaluation criteria (for Service Learning), and justifying textual interpretations (for Mango Street). In the 2009 study, the correlations between human scores for the general and genre-specific rubric were strong, ranging between .75 and .89 (see Table 11). In the 2011 study, after scoring was adjusted to address issues observed in the 2009 dataset, these correlations are weaker, though still moderate to strong: .79 for Ban Ads and .62 for Mango Street (see Table 11).

Table 11

*Correlations between the Common Writing Quality Rubric and Genre-specific Rubrics for Each Prompt in the 2009 and 2011 Studies (p<.001)*



| Prompt | 2009 | | 2011 | |
|---|---|---|---|---|
| Genre-Specific Rubrics | Human Scores: Common Writing Quality Rubric | E-rater Model Trained on Common Writing Quality Rubric | Human Scores: Common Writing Quality Rubric | E-Rater Model Trained on Common Writing Rubric |
| Ban Ads | .86 | .76 | .79 | .78 |
| Mango Street | .80 | .80 | .62 | .61 |
| Service Learning | .75 | .68 | n/a | |

This result suggests an increase in differentiation between the two human rubrics, particularly for the Mango Street prompt. The .64 cross-rubric correlation that we see in the 2011 Mango Street administration is small enough to support the conclusion that the human raters were responding to different traits - consistent with a testing plan in which human scores are used for some aspects of writing quality, while AES systems are used for others.

If the 2011 human scoring distinguishes more reliably between the traits identified in the general rubric and the traits identified in the genre-specific rubrics, then we would also expect a more consistent picture when we compare how human and e-rater scores on the general rubric correlate with human scores on the genre-specific rubric. This is, in fact, the pattern that we observed (see Table 11). In the 2009 data, correlations between scores on the general writing/genre-specific writing rubrics are significantly reduced if we substitute e-rater scores for human general writing scores. This difference is particularly striking for Ban Ads (where the correlation is reduced from .86 to .76, and for Service Learning, where the correlation is reduced from .75 to .68. By contrast, in the 2011 data, the correlations remain similar if we substitute e-rater scores for human general writing quality scores: .79 vs. .78 for Ban Ads, and .62 vs. .61 for Mango Street. Since the e-rater scores were trained on general writing scores from 2009, and cannot reflect changes to the general writing scores in the 2011 data, these results suggests that the increased consistency may be due to changes in the consistency with which the genre-specific rubrics were applied.

***Comparison with lead-in scores.***

Both human generic-rubric and e-rater essay scores were moderately correlated with total score on the lead-in tasks, as shown in Table 12. There were no strong trends or patterns between the 2009 and 2011 administrations. The correlations fall in the range .55 to .61, high enough to confirm that the lead-in tasks are measuring related skills, but not so high as to make them equivalent to essay scores.

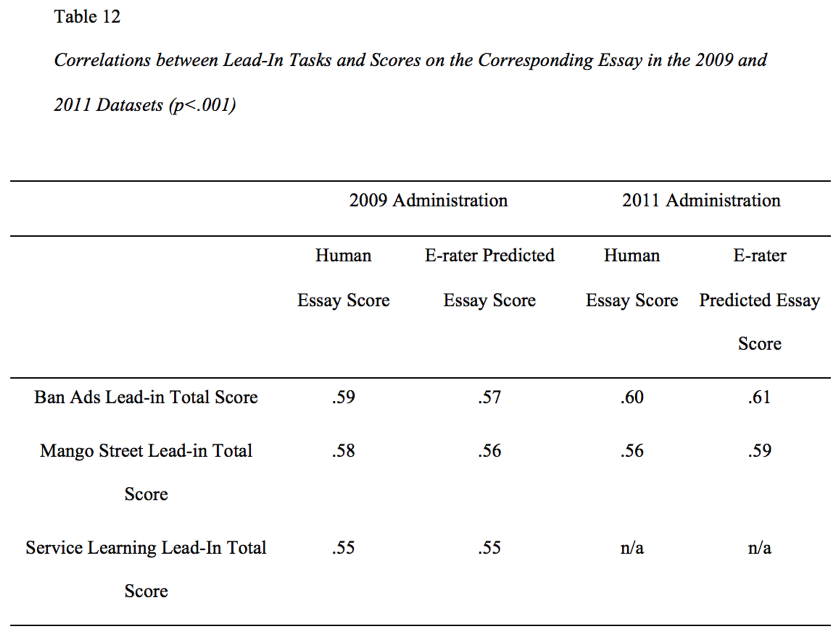


Table 12

*Correlations between Lead-In Tasks and Scores on the Corresponding Essay in the 2009 and 2011 Datasets (p<.001)*

| | 2009 Administration | | 2011 Administration | |
|---|---|---|---|---|
| | Human Essay Score | E-rater Predicted Essay Score | Human Essay Score | E-rater Predicted Essay Score |
| Ban Ads Lead-in Total Score | .59 | .57 | .60 | .61 |
| Mango Street Lead-in Total Score | .58 | .56 | .56 | .59 |
| Service Learning Lead-In Total Score | .55 | .55 | n/a | n/a |

***Comparison with reading test scores.***

Both human generic-rubric and e-rater essay scores were moderately correlated with total scores on the CBAL reading tests, as shown in Table 13. They fall in the range .47 to .60: Once again, the scores are high enough to support the conclusion that related skills are involved, but not so high as to suggest that the reading and writing tests are equivalent. There were slightly higher (and statistically significant, p < .05) correlations between reading and writing tests in the same genre (Mango Street to the 'Seasons' reading test focused on literary texts, Ban Ads to the 'Wind Power' reading test focused on informational texts) as compared with those in different genres (Seasons to Ban Ads, Wind Power to Mango Street).

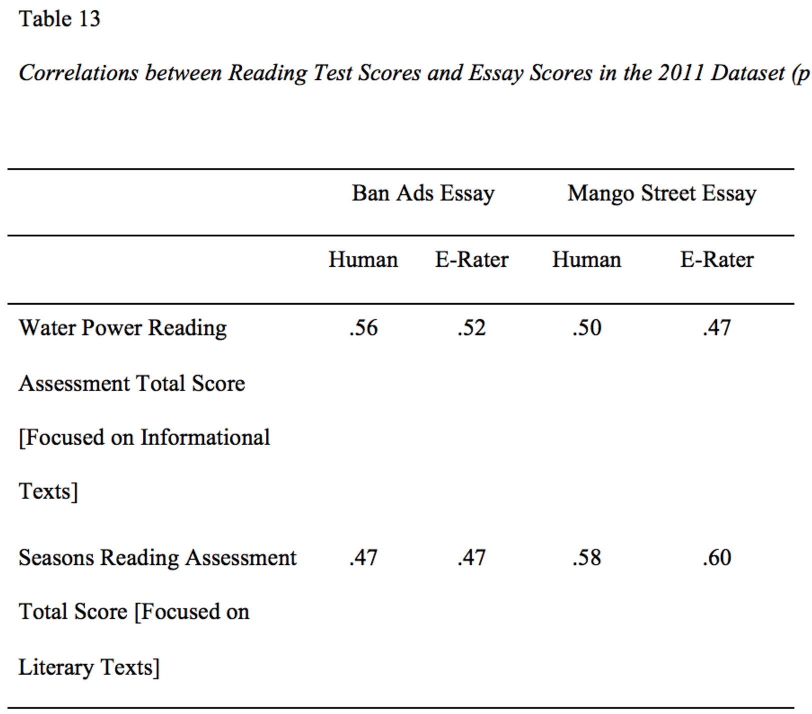


Table 13

*Correlations between Reading Test Scores and Essay Scores in the 2011 Dataset (p<.001)*

| | Ban Ads Essay | | Mango Street Essay | |
|---|---|---|---|---|
| | Human | E-Rater | Human | E-Rater |
| Water Power Reading Assessment Total Score [Focused on Informational Texts] | .56 | .52 | .50 | .47 |
| Seasons Reading Assessment Total Score [Focused on Literary Texts] | .47 | .47 | .58 | .60 |

***Comparison with keystroke log features.***

When document length is controlled for, certain timing features (e.g., the mean latency and duration of pauses within and between words) are negatively correlated with human and e-rater score, while other features (such as the variability of pauses between sentences) are positively correlated with human and e-rater score. This pattern has a straightforward cognitive interpretation, in which pauses between sentences correspond to pauses for planning, and pauses in and between words, to dysfluency in text production, and in fact the literature suggests that more skilled writers, for whom fluency of basic text production is not an issue, devote more time to pauses at likely locations for planning, such as at or near a clause or sentence boundary and less to monitoring character-level text production (Chanquoy, Foulin, & Fayol, 1990; Flower & Hayes, 1980; Kellogg, 2001; Olive & Kellogg, 2002). The fact that the e-rater model produces a similar pattern of correlations to human scores on the 2011 dataset, involving a different sample than that on which it was trained, provides at least preliminary support for the generalizability of the model (see Table 14).

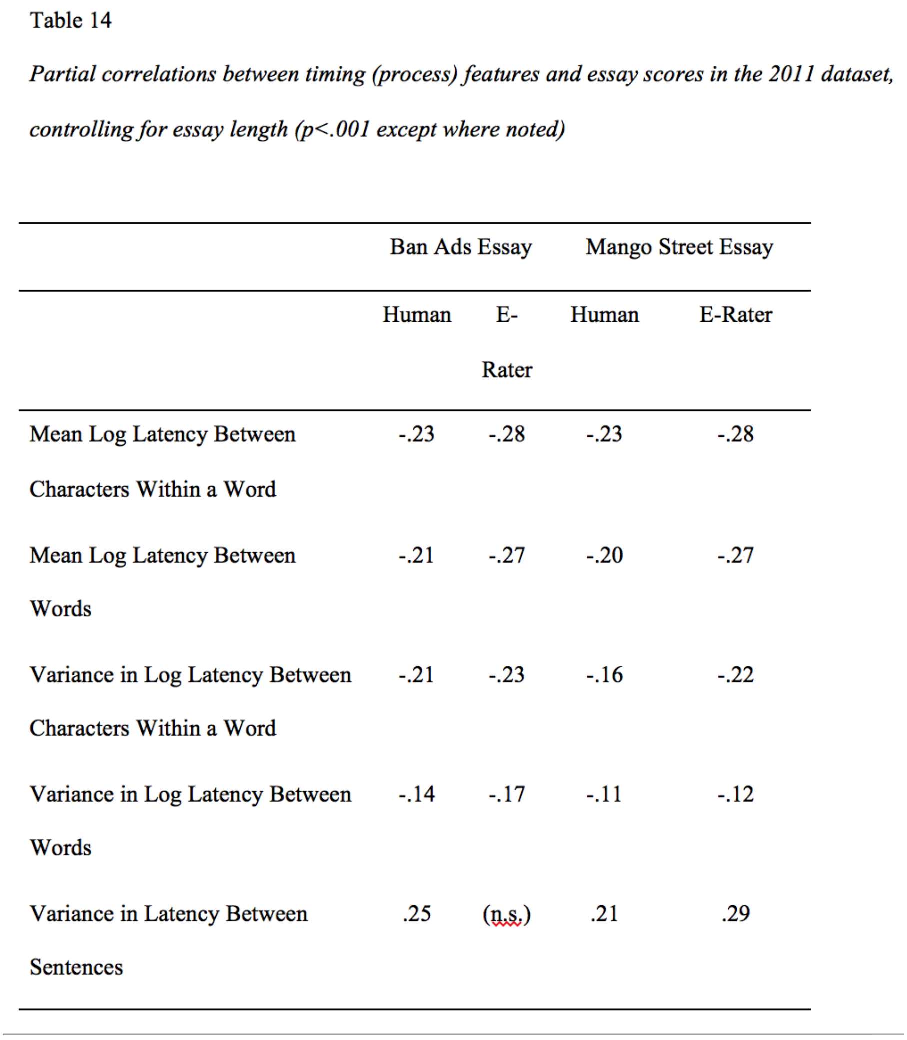


Table 14

*Partial correlations between timing (process) features and essay scores in the 2011 dataset, controlling for essay length (p<.001 except where noted)*

| | Ban Ads Essay | | Mango Street Essay | |
|---|---|---|---|---|
| | Human | E-Rater | Human | E-Rater |
| Mean Log Latency Between Characters Within a Word | -.23 | -.28 | -.23 | -.28 |
| Mean Log Latency Between Words | -.21 | -.27 | -.20 | -.27 |
| Variance in Log Latency Between Characters Within a Word | -.21 | -.23 | -.16 | -.22 |
| Variance in Log Latency Between Words | -.14 | -.17 | -.11 | -.12 |
| Variance in Latency Between Sentences | .25 | (n.s.) | .21 | .29 |

**Usefulness**

***Comparison with total test scores.***

The e-rater model's scores were strongly associated with overall test performance, correlating .78 with total test score for Ban Ads, and .73 with total test score for Mango Street. In fact, if we combine the e-rater models with other automatically scored items (i.e., scores on the selected response, but not constructed response, lead-in questions), we can predict total test score at very high levels of accuracy (R=.91, adjusted $R^2$ = .83 for Ban Ads; R=.85, adjusted $R^2$ = .72 for Mango Street).

## Discussion

The overall pattern of results we have considered thus far can be summarized as follows.

**Accuracy of AES scoring**

The e-rater model stably predicts human scores on the generic writing rubric on which it was trained, even in the 2011 dataset that drew upon a different set of schools, but was targeted at the same general student population. The e-rater models appear to meet the criteria for operational use of an AES model, with one exception, which can be accounted for by adding a feature to detect cases of plagiarism. We can therefore answer the first research question affirmatively: It is possible to train an operational AES model to score writing from sources. In fact, the e-rater model appears to be somewhat more consistent than a single human rater, though the differences are not large.

Some of the complexities we encountered in the model-building process, particularly the differences between the 2009 and 2011 populations, underscore the importance of understanding exactly what the human raters are rating in an essay. It appears to have mattered very much whether human raters were sensitized to the way that writers handled sources, particularly cases of plagiarism. And conversely, it appears to matter whether the AES model includes features that address plagiarism. Including a feature focused on plagiarism in the AES model improved agreement with human scorers on both prompts. In fact, on the 2011 Mango Street administration, the e-rater model performed below expected levels unless plagiarism was taken into account.

**Validity Evidence**

The 2011 Mango Street administration also indicates that the performance of the raters on different rubrics can diverge, as they should if they are focused on different constructs. The .62 correlation between the general writing quality rubric and the genre-specific rubrics is high enough to reflect the relationship between the two rubrics, but not so high that we could simply substitute one for the other (see Table 11). The correlation with the lead-in tasks is also in the moderate range, about what one would expect if the skills measured in the lead-in tasks are related, but not identical, to the features directly measured by e-rater when it scores the writing task. The other measures we examined, including reading test scores and patterns of pauses in keystroke logs, also showed appropriate relationships to the e-rater scores.

In other words, the pattern of results suggests that the e-rater model validly addresses some aspects of writing skill (such as the ability to produce well-structured text fluently without major errors in grammar, usage, mechanics, or style), but not others (such as the ability to justify a literary interpretation.)

**Usefulness**
The e-rater model, by itself, is strongly associated with total test score on the CBAL test designs, and when combined with selected-response scores from the test, predicts 80% of the variance in total test score. This result suggests that e-rater models can validly be used to score essay responses in a test of writing from sources. These results provide a preliminary answer to our third research question.

In fact, the strength of the association between the automatically-scored and the human-scored parts of the test could support a reporting model in which the selected-response and automatically-scored essay portions of the test were used to provide interim score reports for formative purposes, and combined later with human scores.

Note, however, that the strength of prediction is such that scores on the genre-specific rubrics might most usefully be used to identify students who fall off the main trend line. For instance, scores on the genre-specific rubrics would help identify students who produced essays with strong surface traits but which failed to solve the rhetorical problems set by the prompt, or who produced essays that displayed with strong content but which suffered from problems in grammar, usage, mechanics and style. These are critical cases to identify to maintain overall validity of scoring even if they account for a relatively small proportion of the assessed population.

In other words, these results are consistent with the position that automated scoring can be used to score writing from sources, though use of automated scoring may need to be supplemented by additional sources of evidence that provide more direct information about the thinking and reading skills that support effective writing.

**Pattern of Scores across Tasks**
One of the goals of the CBAL test design was to create an assessment in which the pattern of performance across tasks can be interpreted for formative purposes. This aspect of the test design lies outside the scope of the present paper, but it is worth noting that the relationships among tasks generally followed expectations. The hardest task in the Ban Ads task, for instance, was the critique task; the easiest, a task which required students to decide whether specific statements supported banning or allowing advertisements aimed at children. This easy task had an interesting relationship with writing scores: Students who answered fewer than eight of ten questions correctly overwhelmingly scored two or less on the 5-point CBAL writing rubrics. We expect, along with some of our colleagues, to examine the patterns of relationships among lead-in tasks and writing tasks across multiple administrations. Preliminary analyses are presented in Fu, Chung, & Wise (2013), Fu & Wise (2012) and Graf & van Rijn (2013).

## Conclusions

Overall, these results support the strategy adopted in the CBAL research agenda, in which writing skill is measured in multiple ways: with a set of lead-in tasks assessing important contributing skills, by applying multiple rubrics to the essay itself, and by collecting additional evidence about the writing process. When automated scoring is embedded in such an approach, many of the standard criticisms of AES do not apply, because the AES model is not the sole indicator, and other sources of evidence help capture information to which the scoring engine might not be directly sensitive.

In essence, we must start with the recognition that e-rater and other AES systems do not directly provide measurement of the more complex kinds of writing skills in which critical thinking and reading are deeply embedded. But as long as these constructs are measured in some other way - as the lead-in tasks and human-scoring for genre-specific rubrics do for CBAL writing assessments - then the AES engine appears to contribute robust evidence about writing proficiency.

**Author Note**

**Paul D. Deane**, a principal research scientist in Research & Development at Educational Testing Service, earned a Ph.D. in linguistics at the University of Chicago in 1987. He is author of Grammar in Mind and Brain (Mouton de Gruyter, 1994), a study of the interaction of cognitive structures in syntax and semantics. His current research interests include automated essay scoring, vocabulary assessment, and cognitive models of writing skill.

**Frank E. Williams**, an associate research scientist in Research and Development at Educational Testing Service, earned a Ph.D. in Psychometrics at Fordham University in 2013. His research interests include subgroup differences of human and automated scores as well as the use of inferential statistics in quality control procedures for constructed-response items.

**Catherine S. Trapani**, an independent consultant in assessment and education, earned a Ph.D. in Psychometrics from Fordham University in 2013 and an M.S. in Statistics from Montclair State University in 2002. She is co-author of more than twenty articles in educational assessment including automated essay scoring, assessment validity, value-added models and teacher practices. Her current interests include data literacy, high-stakes testing, accountability and adult learners.

**Vincent Z. Weng**, a risk manager in JPMorgan Chase, earned a M.S. in Statistics from Mississippi State University in 2000 with Minor in Computer Science Database Management.His current interests include risk modeling, NLP modeling and data mining.

**Corresponding Author**:
Paul Deane (pdeane@ets.org)
(609) 734-1927

## References

Almond, R., Deane, P., Quinlan, T., & Wagner, M. (in press). *A preliminary analysis of keystroke log data from a timed writing task*. Princeton, NJ: Educational Testing Service.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, D.C.

Attali, Y. (2013). Validity and Reliability of Automated Essay Scoring. In Shermis, M.D. & Burstein, J (Eds.), *iHandbook of automated essay evaluation: Current applications and new directions* (pp. 181-198). New York, N.Y.: Routledge.

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

Bennett, R. (2011). *CBAL: Results from piloting innovative K-12 assessments* (Research Report 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) Berlin, Germany: Springer.

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

Chanquoy, L., Foulin, J.-N., & Fayol, M. (1990). Temporal management of short text writing by children and adults. *Cahiers de Psychologie Cognitive 10,* 513-540. Cline, F. (2012). *2011 CBAL multi-state reading and writing study* (Internal Project Report). Princeton, NJ: Educational Testing Service.

Common Core State Standards Initiative. (2011). *Common core state standards / anchor standards / college and career readiness standards for writing.* Retrieved from http://www.corestandards.org/the-standards/english-language-arts-standards/anchor-standards/college-and-career-readiness-anchor-standards-for-writing/

Deane, P. (2010). *Covering the writing construct: An exploration of automated essay scoring in the context of a cognitively-based approach to writing assessment* (Internal Project Report). Princeton, NJ: Educational Testing Service.

Deane, P. (2011). *Writing assessment and cognition* (Research Report 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 11-01). Princeton, NJ: Educational Testing Service.

Deane, P., Fowles, M., Persky, H., Keller, B., Cooper, P., Morgan, R., Ecker, M. (2010). *CBAL writing 2010 summative project report* (Internal Project Report). 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,* 151-177.

Deane, P., Quinlan, T., & Kostin, I. (2011). *Automated scoring within a developmental, cognitive model of writing proficiency* (Research Report 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 08-55). Princeton, NJ: Educational Testing Service.

Ericsson, P.F., & Haswell, R. Eds. (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Flower, L., & Hayes, J.R. (1980). The dynamics of composing: Making plans and juggling constraints. In L.W. Gregg & E.R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 31-50). Hillsdale, NJ: Erlbaum.

Freedman, S. W. (1984). The registers of student and professional expository writing. Influences on teacher responses. In R. Beach & S. Bridwell (Eds.), *New directions in composition research* (pp. 334-347). New York, NY: Guilford Press.

Fu, J., Chung, S., & Wise, M. (2013). *Statistical report of fall 2009 CBAL writing tests*. ETS Research Memorandum RM-13-01. Princeton, NJ: Educational Testing Service.

Fu, J., & Wise, M. (2012). *Statistical report of 2011 CBAL multistage administration of reading and writing tests*. ETS Research Report RR-12-24. Princeton, NJ: Educational Testing Service.

Graf, E.A., & van Rijn, P. (2013). Recovery of CBAL Learning Progressions: Theory, Results, Challenges, and Next Steps. Paper presented at the annual meeting of the National Conference on Measurement in Education, San Francisco, April 29, 2013.

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

Hillocks, G., Jr. (2008). Writing in secondary schools. In C. Bazerman (Ed.), *Handbook of research on writing: History, society, school, individual, text* (pp. 311-330). Mahwah, NJ: Erlbaum.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60,* 237-263.

Kellogg, R.T. (2001). Competition for working memory among writing processes. *American Journal of Psychology, 114,* 175-192.

Lee, Y.-W., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL CBT essays: Scores from humans and e-rater®* (TOEFL Research Report 08-01). Princeton, NJ: Educational Testing Service.

McCutchen, D. (2006). Cognitive factors in the development of children's writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 115-130). New York, NY: Guilford.

Miller, K., Lindgren, E., & Sullivan, K. (2008). The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *TESOL Quarterly, 42,* 433-454.

Murphy, S., & Yancey, K.B. (2008). Construct and consequence: Validity in writing assessment. In C. Bazerman (Ed.), *Handbook of research on writing: History, society, school, individual, text* (pp. 365-386). New York, NY: Erlbaum.

Olive, T. & Kellogg, R.T. (2002). Concurrent activation of high- and low-level production processes in written composition. *Memory and Cognition, 30,* 594-600.

PARCC. (2011). The Partnership for Assessment of Readiness for College and Careers (PARCC) Application for the Race to the Top Comprehensive Assessment Systems Competition. Retrieved from http://www.parcconline.org/sites/parcc/files/PARCC%20Application%20-%20FINAL.pdf

Perin, D. (2009). 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.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards: The New US Intended Curriculum. *Educational Researcher, 40,* 103-116.

Ramineni, C., Williamson, D.M. (2013). Automated essay scoring: Psychometric guidelines and practices, *Assessing Writing, 18,*(1), 25-39.

Shermis, M. D. & Burstein, J. (Eds.) (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, N.J.: Lawrence Erlbaum.

Shermis, M., Burstein, J., & Leacock, C. (2006). Applications of computers in assessment and analysis of writing. In C.A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 403-417). New York, NY: The Guilford Press.

Shermis, M., & Hamner, B. (2012, April). Contrasting state-of-the-art automated scoring of essays: Analysis. In Shermis, M.D. & Burstein, J (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 323-346). New York, N.Y.: Routledge.

Smarter Balanced Assessment Consortium. (2012a). *Smarter Balanced English Language Arts item and task specifications*. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/EnglishLanguageArtsLiteracy/ELAGeneralItemandTaskSpecifications.pdf

Smarter Balanced Assessment Consortium. (2012b). *SBAC Appendices*. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/SBAC_Appendices.pdf

State of Florida Department of Education. (2012). *Invitation to negotiation: PARCC item development*. Retrieved from http://myflorida.com/apps/vbs/vbs_www.ad.view_ad?advertisement_key_num=98159

Wengelin, A., Torrance, M., Holmqvist, K., Simpson, S., Galbraith, D., Johansson, V., & Johansson, R. (2009). Combined eye-tracking and keystroke-logging methods for studying cognitive processes in text production. *Behavior Research Methods, 41,* 337-351.

Williams, F.E., & Breyer, F. J. (2011). *Empirically evaluating rater behavior by using e-rater as an anchor*. Unpublished manuscript.

Williamson, D., Xi, X., Breyer, F.J. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*, 2-13.