Research Report

Examining the Writing Processes in Scenario-Based Assessment Using Regression Trees

ETS RR–20-18

Yi Cao
Jianshen Chen
Mo Zhang
Chen Li

December 2020

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

Examining the Writing Processes in Scenario-Based Assessment Using Regression Trees

Yi Cao, Jianshen Chen, Mo Zhang, & Chen Li

Educational Testing Service, Princeton, NJ

Scenario-based writing assessment has two salient characteristics by design: a lead-in/essay scaffolding structure and a unified scenario/topic throughout. In this study, we examine whether the scenario-based assessment design would impact students' essay scores compared to its alternative conditions, which intentionally broke the scaffolding effect and/or topic effect. Furthermore, with the availability of rich keystroke log data, we used tree-based methods to investigate which writing process features had greater influence in predicting students' essay scores under different assessment conditions. The results revealed significantly lower essay scores when the unified topic effect was removed and no significant essay score change when neither scaffolding nor topic effects existed compared to the original scenario-based assessment condition. The findings also suggested that different assessment conditions call on different configurations of writing process features. The rate of typo correction in the writing process was consistently among the highest ranking features that predicted the essay scores well, regardless of the presence of scaffolding and/or topic effects. The broken topic effect might affect the extent of word finding and retrieval as well as editing across multiple words by the students during writing. Topics for future study and limitations are also discussed.

Keywords CBAL; scenario-based assessment; keystroke logging; writing process; regression trees

doi:10.1002/ets2.12301

Recent theoretical advances in the competency model and learning progressions lay the groundwork for the development of a scenario-based assessment. The competency model describes knowledge, processes, strategies, and practices central to proficiency in a certain domain as well as how those competencies might be interrelated. The learning progressions map out the pathways that most students are hypothesized to take toward mastery of certain critical competencies in the competency model. Rooted in these advances, a scenario-based assessment intends to serve dual purposes to provide valid measurement of student competencies and to better inform and support learning and teaching practice (Bennett et al., 2016). Scenario-based assessment design has been applied to a variety of areas, such as reading (e.g., O'Reilly & Sabatini, 2013), mathematics (e.g., Cayton-Hodges et al., 2012), and science (e.g., Quellmalz et al., 2013). The focus of the present study is on scenario-based assessment of writing (e.g., Bennett et al., 2016; Deane et al., 2011) and, in particular, how this assessment design may affect students' writing processes.

In the scenario-based writing assessment, items are written on the same topic or scenario and are sequenced based on a theoretically determined order. The assessment first requires students to complete a series of lead-in tasks on a unified topic and then asks the students to compose a culminating essay on the same topic. Two salient characteristics are embedded in this assessment. First, this assessment, by design, has a scaffolding structure; that is, lead-in tasks are followed by a culminating essay writing task. Second, this assessment uses a unified scenario throughout. To examine the potential impacts of the scaffolding structure and the unified topic on a student's writing proficiency estimation, Zhang, van Rijn, et al. (2019) conducted an experimental study in which they created alternative forms to intentionally break these two effects. They found that the theoretically motivated scenario-based design did not appear to artificially increase total-test or essay scores, and it functioned as well as, and sometimes better than, the alternative forms in terms of the psychometric characteristics examined, such as total test score reliabilities.

With the availability of rich keystroke log data collected during the essay text production, in this exploratory study, we examined and compared the essay writing processes in the original scenario-based assessment form and its alternative forms from a new perspective. Previous studies have tackled similar problems. For example, Zhang et al. (2017) compared the scenario-based assessment form with theoretically sequenced items to an alternative form that broke the item

Corresponding author: Y. Cao, E-mail: ycao@ets.org

sequencing. They found that students spent more time on text editing and revising when given support on task and topic preparation. Guo et al. (2019) also found that the placement of lead-in tasks in the scenario-based assessment prior to the essay task enabled students to produce essays similar in quality but with less time and using fewer words. A larger suite of process features became available after Zhang and colleagues investigated only the scaffolding effect. They did not examine the topic effect in their study. Guo and colleagues examined both the topic and scaffolding effects from a stochastic writing process perspective in which four writing states were identified and examined (i.e., text production, long pause, editing, and jump editing) among the test forms.

As a complement to the previous studies, it is worthwhile to investigate how the current large suite of process features may differ in scenario-based assessment form compared to its alternative forms. Sinharay et al. (2019) noted that data mining methods would be suitable for dealing with high dimensional data. Hence, in this study, we used data mining methods, in particular, regression trees, to investigate the driving factors that could help explain assessment form differences and also to identify writing process features that can best differentiate students' writing performance for each assessment form.

The organization of the report is as follows. We first provide background information on the scenario-based summative assessment on argumentative writing as well as an overview of keystroke logging. Then we propose two research questions to investigate. Next, we elaborate the methodology of this study, including assessment forms and conditions, participants, variables, and analyses. We then present the study results. Finally, we summarize the findings and discuss the limitations of the study, future directions, and the implications of this study for future writing studies.

Scenario-Based Summative Assessment of Argumentative Writing

Argumentation skills (reading, writing, and critical thinking) are the core skills in *discuss and debate ideas*, one of the 11 key practices in the English language arts competency model developed through the *CBAL*® learning and assessment tool research initiative (see, for a detailed overview of the 11 key practices, Bennett, 2010; Bennett et al., 2016; Deane et al., 2015). Deane and Song (2015) did an extensive literature review and categorized argumentation into five distinct but intertwined phases of core activities (and related sets of targeted skills) to understand the concept, including *context and stakes* (appeal building skills), *explore the subject* (research and inquiry skills), *consider positions* (taking a position skills), *create and evaluate arguments* (reasons and evidence skills), and *organize and present arguments* (framing a case skills). Among them, four phases, except for *explore the subject*, are unique to the argumentation, and therefore four progress variables, each related to a different phase, are defined and linked to the major developmental stages or learning progression levels. The detailed evidence models are further defined for each progress variable in terms of reading, writing, and critical thinking skills, and corresponding tasks are mapped onto different levels of learning progressions. Interested readers may consult Bennett et al. (2016) and Deane and Song (2015) for more elaborate descriptions of the competency model and corresponding learning progressions on argumentation.

Guided by all the preceding extensive research, a specific scenario-based summative assessment on argumentative writing was created (Deane et al., 2018). In this assessment, students are presented with a scenario describing a debatable issue then asked to complete a sequence of tasks, read and summarize arguments from multiple source materials, critique other people's arguments, and classify arguments and claims/evidence from both sides of the issue. Finally, students are required to present their own arguments with evidence and reasoning in an essay using the same source materials. While assessing certain subskills in argumentation and writing, the lead-in tasks are designed to help students get familiar with a topical context and source materials and to guide them through steps in an expert writing process so that they can write more effectively and efficiently when they undertake the culminating essay task.

Keystroke Logging

Keystroke logging has become one of the established research methods for writing research (Sullivan & Lindgren, 2006). As an observational tool, keystroke logging involves a nonobtrusive, real-time recording of all mechanical operations (e.g., key presses and mouse clicks or movements) and their associated temporal information (e.g., key in and key out time, pause length) as writers compose on the computer (Leijten & Van Waes, 2013). The output log file stores highly detailed records that include several kinds of information, such as action, duration, location, and time point (Zhang et al., 2017; Zhang & Deane, 2015). For example, a writer typed "Engligh," deleted "g," and added "s" as an attempt at spelling error

correction. This kind of action sequence, along with the time stamp for each key press, is precisely recorded in a keystroke log and can be analyzed for various purposes. Different keystroke logging programs have been developed and are available online, including jEdit (1998), ScriptLog (2006), Inputlog (2013), and Translog (2006).

Keystroke logging allows not only for the recording and storage of information related to writing activities but also for the subsequent extraction of features to characterize a writing process. The process features considered in this study are extracted from the Educational Testing Service (ETS) keystroke logging program (Deane et al., 2016). Among them, some could be viewed as indicators of fluency, such as the median pause time between keystrokes and the median value of the longest within-word pause. Some others measure the extent of editing and revision, such as the proportion of characters deleted in the process of eliminating multiple words as a function of the total number of inserted and deleted characters or the proportion of words that are subjected to minor editing (one- or two-character changes) as a function of the total number of keystroke records (including characters that were later deleted). Still others measure the extent of text or idea planning and deliberation, such as the proportion of time spent on long pauses occurring at the beginning of a sentence, a string of fluent text production as a function of total writing time, or the median pause length at sentence junctures (Zhang & Deane, 2015).

Keystroke logging has been applied to the areas of the linguistic, textual, and cognitive studies of writing as well as to language learning and pedagogy (Sullivan & Lindgren, 2006). However, it has not been researched extensively in the context of educational assessment (e.g., Deane, 2014; Deane & Zhang, 2015; Sinharay et al., 2019; Zhang, Bennett, et al., 2019). How writing processes differ under different assessment conditions in the context of educational assessment was of particular interest in the present study. In particular, two major research questions were addressed:

Research Question 1: Does the scenario-based assessment design introduce a scaffolding effect and/or topic effect that impacts students' essay scores?

Research Question 2: For each assessment condition, what are the strongest predictors of essay scores among the writing process features and demographic variables? Of the strongest predictors, what are the main driving factors that differentiate writing performance?

Method

Assessment Forms and Conditions

We used a data set collected by Zhang, van Rijn, et al. (2019). Four versions of a scenario-based assessment on argumentative writing were administered in that study. Each form had four tasks, including three lead-in tasks (Tasks 1, 2, and 3) and one essay writing task (Task 4). Table 1 provides descriptions of the four forms.

Form 1 in this study was the original scenario-based assessment form (referred to as the base form), which started with three lead-in tasks that required students to read, think, and respond to questions related to several source documents. Task 4 was the essay. The scenario throughout Form 1 was about whether the United States should ban advertising to children under age 12 years (denoted as BAN). Form 1 had both the intended scaffolding structure (i.e., lead-in tasks followed by an essay writing) and the unified topic throughout (i.e., lead-in tasks and essay have the same topic). The assessment as a whole is designed to be administered in two test sessions, for which each session is one class period time. Also, note that Tasks 1 and 2 were designed to be given in one test session, and Tasks 3 and 4 were designed to be given in a second test session.

Three alternative forms intentionally broke the two characteristics of the scenario-based assessment: the scaffolding structure and/or the unified topic. Form 2 offered exactly the same four tasks as Form 1 but presented the essay task (Task 4) first, so that the scaffolding structure was broken. It should be noted that the essay in Form 2, since administered first, was introduced with the setting, purpose, and three source documents.

Form 3 kept the same task administration order as Form 1 (lead-in tasks first and then essay writing) but changed the scenario in lead-in tasks. The lead-in tasks in Form 3 were taken from a parallel test form on the topic of whether schools pay students for getting good grades (denoted as CFG).[1] Task 4 still required students to write an essay on the same topic as Form 1 (with the same setting, purpose, and three source documents from Form 1 BAN). Form 3 was intended to have lead-in tasks on one scenario and the concluding essay on another scenario.

Form 4 was offered in the reverse order of Form 3, which presented the essay writing first with the same setting, purpose, and three source documents from Form 1 BAN, followed by the lead-in tasks as used in Form 3 CFG. In that way, Form 4 broke the scaffolding structure as well as had different topics between the essay writing and lead-in tasks.

Table 1 Description of the Four Assessment Forms

Form	Session 1	Session 2	Scaffolding structure presence?	Unified topic throughout?
1	BAN Task 1, BAN Task 2	BAN Task 3, BAN Task 4	Yes	Yes
2	BAN Task 4, BAN Task 3	BAN Task 1, BAN Task 2	No	Yes
3	CFG Task 1, CFG Task 2	CFG Task 3, BAN Task 4	Yes	No
4	BAN Task 4, CFG Task 3	CFG Task 1, CFG Task 2	No	No

Note. Form 1 is the base form. Scenarios are Ban Advertising for Children (BAN) and Providing Cash for Grades (CFG). Task 1 is about "read and summarize arguments"; Task 2 is about "evaluate argument"; Task 3 is about "analyze arguments"; Task 4 is about "present your view in an essay." Tasks 1 – 3 are lead-in tasks, and Task 4 is the essay of interest.

Table 2 Description of the Three Assessment Conditions

Condition	Corresponding form	Scaffolding effect on essay writing?	Topic effect on essay writing?
A	Form 1	Yes	Yes
B	Form 3	Yes	No
C	Forms 2 and 4	No	N/A

Note. Condition A is the base condition.

Because the focus of our study was the impact of writing process features and demographic variables in essay writing, we further collapsed two of the four assessment forms (see Table 2) and ran analyses based on these three conditions: Condition A (Form 1) was the base condition with both scaffolding and topic effects on essay scores, Condition B (Form 3) included the scaffolding effect but not the topic effect on essay writings, and Condition C (Forms 2 and 4) had neither scaffolding nor topic effects. We combined Forms 2 and 4 into Condition C because both forms presented the essay writing first, followed by lead-in tasks, regardless of whether there was a unified topic throughout. Conceptually, writing process features and demographic variables would relate to students' essay scores to the same degree on Forms 2 and 4.

Three lead-in tasks contributed to a maximum of 28 score points, and the essay was scored from 0 to 10 using two rubrics. Rubric 1 focused on writing fundamentals (e.g., word usage, writing mechanics, vocabulary, organization, and development on a 6-point scale ranging from 0 to 5), and Rubric 2 focused on the high-level thinking specific to argumentative writing (e.g., the quality of idea, strength of argument, and factual accuracy, also on a 6-point scale ranging from 0 to 5). A human score of 0 was given to essays with special characteristics, such as empty, off-topic, plagiarized, or random keystrokes. Those essays were removed from our analyses.

Participants

Data were collected over a 1-month period in 2014 and included a sample of 1,050 eighth-grade students in the United States. Students were randomly assigned one of the four forms within each participating class. Each form was administered in two separate 45-min class sessions on different days in the same week.

The keystroke process features would be less reliable if the total writing time/session were too short. For this reason, we further removed responses that were too short in terms of the writing time or length. Specifically, we followed data cleaning steps before conducting further analyses. A student's record was excluded if (a) the keystroke log was corrupted due to unexpected technological glitches, (b) the essay contained fewer than 25 words, (c) the recorded total time spent on the essay was less than 3 min or more than 35 min, (d) the essay was scored as 0 by human raters, or (e) the student's demographic information was missing. The final analysis data set included 846 records; the descriptive statistics for essay scores by each demographic variable are provided in Table 3.

Variables

Three outcome variables were analyzed: essay score using Rubric 1; essay score using Rubric 2; and total essay score, which is the sum of Rubric 1 and Rubric 2 scores. Two sets of variables were used as covariates/predictors: (a) five demographic variables — gender, ethnicity, free/reduced-price lunch as socioeconomic status, English language learner status, and Title 1 accommodation (Table 3 provides the descriptive statistics by each of the five demographic variables), and (b) 34 writing process features extracted from keystroke logs (the keystroke logging process features are described in Table 4).

Table 3 Descriptive Statistics for Essay Scores by Demographic Variable

Group	*N*	%	Score, M (*SD*)		
			Rubric 1	Rubric 2	Total essay
Gender					
Female	447	52.84	2.24 (0.97)	2.64 (1.00)	4.88 (1.77)
Male	399	47.16	2.00 (0.85)	2.36 (0.91)	4.37 (1.58)
Ethnicity					
African American	75	8.87	1.61 (0.70)	1.99 (0.83)	3.60 (1.38)
Asian	22	2.60	2.27 (0.77)	2.86 (0.89)	5.14 (1.39)
Hawaiian/Pacific Islander	3	0.35	–	–	–
Hispanic	33	3.90	1.91 (0.80)	2.18 (0.81)	4.09 (1.47)
Middle Eastern	1	0.12	–	–	–
Mixed race	8	0.95	2.00 (1.07)	2.25 (0.71)	4.25 (1.04)
Native American	6	0.71	1.33 (0.52)	1.83 (0.75)	3.17 (0.75)
White	698	82.51	2.19 (0.93)	2.57 (0.97)	4.76 (1.70)
Free/reduced-price lunch[a]					
No	637	75.30	2.21 (0.92)	2.59 (0.99)	4.80 (1.71)
Yes	209	24.70	1.87 (0.89)	2.28 (0.87)	4.15 (1.58)
ELL status					
Current	144	17.02	2.39 (0.92)	2.67 (0.94)	5.06 (1.62)
Former, reclassified as proficient	8	0.95	2.13 (0.83)	2.38 (0.92)	4.50 (1.51)
Initially proficient at school entry	694	82.03	2.07 (0.91)	2.48 (0.98)	4.55 (1.71)
Title 1 accommodation					
No	825	97.52	2.14 (0.92)	2.52 (0.98)	4.66 (1.71)
Yes	21	2.48	1.57 (0.75)	2.24 (0.70)	3.81 (1.25)

Note. Means and standard deviations from groups with $N \leq 5$ are not included. ELL = English language learner. [a] Used as a measure of socioeconomic status.

Data Analysis

Research Question 1 has been partially addressed in Zhang, van Rijn, et al. (2019). In their studies, they used one-way analysis of variance (ANOVA) to examine total assessment score mean differences and total essay score mean differences across forms. For the specific purpose of this study, we analyze only essay scores, but for each rubric and for the total scores. Furthermore, our analyses were based on three assessment conditions, whereas theirs focused on four assessment forms. We ran one-way ANOVA to test the null hypothesis that the means of essay scores are equal across all three conditions. This was followed by Dunnett's method for multiple comparisons between base Condition A and alternative Condition B or C. Dunnett's method is a planned, pairwise procedure that gives adequate family-wise error rate protection. Specifically, mean differences between Condition A and Condition B would indicate the topic effect, and mean differences between Condition A and Condition C would show the scaffolding effect. Three ANOVAs were run separately for Rubric 1 score, Rubric 2 score, and total essay score. For all statistical tests, the critical level was set to be .05.

To answer Research Question 2, we used boosted regression trees (Friedman, 2001; Ridgeway, 1999) to search through all the writing process features and demographic variables to identify key predictors that could best predict essay scores separately for each assessment condition. The R package gbm (Greenwell et al., 2019) was used to conduct boosted regressions. For each boosted regression, we fit 1,000 trees (i.e., run 1,000 iterations) as the suggested rule-of-thumb minimum by Elith et al. (2008) and performed a fivefold cross-validation. In addition to the main effects, up to two-way interactions among predictors were allowed. Furthermore, we set the shrinkage parameter to .05, which represents the learning rate or step-size reduction. The learning rate usually ranges between .001 and .1, and a smaller learning rate leads to slower learning and usually requires more trees. Half the training set observations were randomly selected to propose the next tree in the expansion, which is the default setting of the gbm function of the gbm package. The boosted regression tree analysis was conducted separately for each score (i.e., Rubric 1, Rubric 2, and total essay score) and separately for three assessment conditions. The results of the 1,000 iterations of boosted regression allowed us to obtain a relative influence measure for each predictor. The measures were based on the number of times a variable was selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Elith et al., 2008; Friedman

Table 4 Keystroke Logging Process Features

Process feature	Description
Burst length	Average burst length in words
Characters in multiword deletion	Extent to which deletion of characters occurs in the process of deleting multiple words
Corrected typos	Extent to which correction of mistyped words occurs
Deleted characters	Extent to which deletions of characters occur
Discarded text	Extent to which deletion occurs, as a function of total number of characters in the final submission
Edited chunk	Extent to which deleted text is replaced with edited text of similar content
Edited words	Extent to which words are edited as a function of number of words inserted
End-sentence pause	Median pause time at sentence junctures
Event after last character	Extent to which editing of any kind occurs
In-sentence pause	Extent to which pauses occur at a within-sentence punctuation mark
Interkey pause	Median pause duration between keystrokes
Long jump	Extent to which jumps to a different part of the text occur
Major edit	Extent to which words are edited to make more than a one- or two-character change
Minor edit	Extent to which words are edited to make less than a one- or two-character change
Multiword deletion	Extent to which multiword deletion occurs
Multiword edit time	Extent of time spent in deleting multiple words
New content	Extent to which deleted text is replaced with editing text with different content
Phrasal burst	Extent to which long strings of fluent text production are produced without interruption
Prejump pause	Extent to which a pause occurs just before jumping to a different part of the text
Retyped chunk	Extent to which deleted text is replaced with essentially the same text
Start time	Extent of time spent pausing before beginning writing
Time at burst	Extent to which pauses occur at the beginning of a string of fluent text production
Time between burst	Extent to which pauses occur between strings of fluent text production
Typing speed	In-word typing speed
Typo corrected chunk	Extent to which text is replaced with edited text that differs only in minor spelling correction
Typo correction rate	Extent to which typos were corrected as a function of number of words inserted
Uncorrected spelling errors	Extent to which a spelling error occurs that is not corrected before another unrelated action is taken
Word choice	Extent to which words are edited to produce completely different words
Word choice event pause	Extent to which a pause occurs when replacing words with different words
Word edit pause	Extent to which pauses occur within words during text editing
Word-final pause	Median duration of the pauses occurring before typing the last character in a word
Word-initial pause	Median duration of the pauses occurring before typing the first character in a word
Word-internal pause	Median duration of longest within-word pause
Word space pause	Median duration of the pauses occurring before the space that separates two words

& Meulman, 2003). The sum of the relative influence across all predictors equals 100%. We then ranked all predictors based on their relative influence and considered those with a relative influence equal to or greater than 5% as key predictors. There are no fixed rules to determine the number of most important predictors. We made a judgment call to use 5%, which is two times the average influence if evenly distributed.

As a next step, we ran a single regression tree using only those key predictors identified for further examination. The purpose was to assess, of those key predictors, what were the driving factors that differentiated students' writing performance. A similar two-step method was implemented in Chen and Keller (2019). The rpart package was used to conduct single regression tree analyses (Therneau et al., 2019) in R (R Core Team, 2015). We set the complexity parameter, the minimum R^2 increase at each split, at .01, which is our acceptable level of R^2 increase and also the default setting of the rpart package. The maximum layer of trees was set at 15 to allow for a more detailed story to be described by the key predictors, but it should be noted that the bottom layers of the trees are not as robust as top layers. As with the boosted regressions, the single regression tree analysis was conducted separately for each score (i.e., Rubric 1, Rubric 2, and total essay score) and separately for each assessment condition (i.e., Conditions A, B, and C).

Table 5 Analysis of Variance Comparing Observed Scores on Alternative Conditions Against Base Condition (Rubric 1 Score)

Score	*df*	Sum of squares	Mean square	*F*	*p*
Between groups	2	49.72	24.86	31.35	<.0001
Within groups	843	668.49	0.79		
Total	845	718.21			

	Multiple comparisons: Dunnett's *t* (two-sided)		
	Mean difference	95% CI lower bound	95% CI upper bound
Conditions B – A	−.49[a]	−.68	−.29
Conditions C – A	.11	−.05	.28

[a] Comparisons are significant at the .05 level.

Table 6 Analysis of Variance Comparing Observed Scores on Alternative Conditions Against Base Condition (Rubric 2 Score)

Score	*df*	Sum of squares	Mean square	*F*	*p*
Between groups	2	19.19	9.59	10.42	<.0001
Within groups	843	776.22	0.92		
Total	845	795.40			

	Multiple comparisons: Dunnett's *t* (two-sided)		
	Mean difference	95% CI lower bound	95% CI upper bound
Conditions B – A	−.30[a]	−.51	−.09
Conditions C – A	.07	−.11	.25

[a] Comparisons are significant at the .05 level.

Table 7 Analysis of Variance Comparing Observed Scores on Alternative Conditions Against Base Condition (Total Essay Score)

Score	*df*	Sum of squares	Mean square	*F*	*p*
Between groups	2	130.68	65.34	23.80	<.0001
Within groups	843	2,314.64	2.75		
Total	845	2,445.32			

	Multiple comparisons: Dunnett's *t* (two-sided)		
	Mean difference	95% CI lower bound	95% CI upper bound
Conditions B – A	−0.79[a]	−1.15	−0.43
Conditions C – A	0.18	−0.12	0.49

[a] Comparisons are significant at the .05 level.

Results

Results for Research Question 1

To test the omnibus null hypotheses that the means of essay scores across the three assessment conditions were equal, one-way ANOVAs were run separately for Rubric 1 score, Rubric 2 score, and total essay score. The results are summarized in the upper parts of Tables 5 – 7. The one-way ANOVA results showed that overall, observed mean differences among three conditions were significantly different on Rubric 1 score, $F(2, 843) = 31.35$, $p < .0001$; Rubric 2 score, $F(2, 843) = 10.42$, $p < .0001$; and total essay score, $F(2, 843) = 23.80$, $p < .0001$.

Following the rejections of omnibus null hypotheses, a multiple comparison procedure called Dunnett's two-sided *t*-test was conducted for each essay score, and the Type I experiment-wise error was controlled at .05 for planned, pairwise comparisons of two alternative conditions against the base condition. The multiple comparison results are summarized in the bottom portions of Tables 5 – 7. Results showed that, for all three essay scores, there were statistically significant

Table 8 Prediction Accuracy in Essay Scores by Condition

Score	Condition	R^2	Adjusted R^2
Rubric 1 score	A	.61	.52
	B	.49	.37
	C	.55	.51
Rubric 2 score	A	.42	.29
	B	.43	.29
	C	.37	.31
Total score	A	.60	.51
	B	.53	.42
	C	.55	.50

mean differences between Conditions B and A. However, when comparing Condition C with Condition A, no statistically significant mean difference was found. These findings were consistent with those of Zhang, van Rijn, et al. (2019), where total essay score was a dependent variable.

However, lack of statistically significant differences between mean essay scores does not necessarily mean lack of scaffolding and/or topic effects. As reported and discussed in Zhang, van Rijn, et al. (2019), Zhang et al. (2017), and Guo et al. (2019), even though some mean score differences were not statistically different, the writing processes that led to the final essays indeed differed between students taking different forms, because how items were sequenced and the form structure would affect the cognitive load required to complete the writing task. Research Question 2 further addressed this issue in more detail.

Results for Research Question 2

All five demographic variables and 34 process features (39 in total) were included in the analyses. We first examined the R^2 (proportion of the score variation explained by the model) and adjusted R^2 (a modified version of R^2 that has been adjusted for the number of predictors in the model) from boosted regression trees to evaluate how the predictors explained the variation in the essay scores. These statistics for Rubric 1 score, Rubric 2 score, and total essay score on three assessment conditions are summarized in Table 8. Note that the prediction accuracy of Rubric 1 score was better than the prediction accuracy of Rubric 2 score using process features and demographic variables, consistently for all three conditions. For example, the R^2 for Rubric 1 in Condition A was .61, but it was only .42 for Rubric 2. This result was expected and has been reported in other studies (e.g., Deane & Zhang, 2015). The process features measure writing fluency and productivity, which align better with the basic/standard English writing skills evaluated in Rubric 1.

Boosted regression trees were run to rank the relative importance of all 39 predictors. Predictors that had relative influence equal to or greater than 5% are reported. Table 9 summarizes the significance of the top predictors on the three essay scores by assessment conditions. Note that the process feature typo correction rate was consistently among the highest ranking predictive features across all three conditions, with the only exception being Rubric 2 in Condition B. This result was consistent with previous analyses by Sinharay et al. (2019) and Li et al. (2016), which reported that this process feature was highly predictive of essay quality. This pattern was found across assessment conditions, indicating that the extent of making corrections to typos during the writing process had similar association to final essay quality regardless of the presence of scaffolding structure and/or topic effects.

However, another process feature, uncorrected spelling error, showed a somewhat different pattern: It ranked high for Conditions A and C but not for Condition B, in which the topic effect was broken. On the other hand, in contrast to the other two conditions, Condition B revealed word-initial pause and multiword edit time as the top-ranking predictors of essay scores consistently for the two rubric scores and total essay score. On the basis of this finding, we speculated that the sudden shift of topic in the essay task might have led to more cognitive stress to the students such that they placed more emphasis on word finding, retrieval, and text editing across multiple words, even though the same source materials were supplied. It would be worthwhile as a follow-up study to investigate whether pauses, particularly longer pauses associated with word initials, are topic specific. Similarly, another valuable follow-up study would be to further investigate whether edited words or words left with uncorrected spelling errors interact with topic.

Table 9 Relative Importance of Predictors on Essay Scores by Condition

	Rubric 1 score		Rubric 2 score		Total essay score	
Condition	Predictor	Relative influence (%)	Predictor	Relative influence (%)	Predictor	Relative influence (%)
Condition A	Typo correction rate	12.46	Typo correction rate	7.95	Typo correction rate	11.57
	Uncorrected spelling errors	12.08	Uncorrected spelling errors	7.50	Uncorrected spelling errors	11.54
	Phrasal burst	8.90	Characters in multiword deletion	5.44	Characters in multiword deletion	8.03
	Characters in multiword deletion	7.62	Start time	5.29	Phrasal burst	5.98
Condition B	Word-initial pause	13.28	Word-initial pause	10.76	Word-initial pause	16.71
	Typo correction rate	8.91	End-sentence pause	9.13	Prejump pause	6.57
	Multiword edit time	5.34	Time at burst	6.66	End-sentence pause	6.36
	Burst length	5.17	Multiword edit time	5.55	Typo correction rate	6.23
					Multiword edit time	6.20
					Word-final pause	5.08
Condition C	Typo correction rate	11.87	Typo correction rate	8.47	Typo correction rate	10.82
	Uncorrected spelling errors	8.69	Word choice	8.43	Uncorrected spelling errors	9.06
			Uncorrected spelling errors	7.35		
			Word space pause	5.25		

Note. Predictors that have relative influence >5% were chosen and used to fit regression trees.

Another finding worth noting was that Condition B results also indicated the importance of the end-sentence pause and prejump pause features, for which pattern was largely absent for the other two conditions. This result suggested that the quality of the final response in Condition B tended to rely more on phrase and sentence-level editing and planning when the topic effect was removed. Additionally, while Condition A placed more emphasis on text production fluency and word-level editing and revision, as evidenced by the features phrasal burst and characters in multiword deletion ranking high from the boosted regression trees, breaking the unified topic (in Condition B) might have shifted the emphasis to word finding and retrieval (indicated by the highest ranking of the feature word-initial pause) as well as editing. It is possible that students' lower familiarity with the topic in Condition B (which was further exacerbated by the shift in topic in the essay task), compared to students in Condition A, led to more cognitive difficulties in topic-specific vocabulary retrieval and editing involving more than a single word. Finally, unlike Conditions A and B, in which several writing process features had relative influence measures greater than 5%, in Condition C, the driving factors were largely typo and spelling error correction-related behaviors. Only these two features were listed as top predictors in Table 9 for Rubric 1 and total essay score. Two more features were added in Rubric 2: word choice and work space pause. This result indicated that when both the scaffolding and topic effects were removed in an assessment (Condition C), the writing performance was primarily driven by error correction-related writing behaviors.

Subsequently, predictors identified in Table 9 were used to fit single regression trees to further examine how these top predicting features differentiated students with different writing performance. Single regression trees were plotted for easier visualization and interpretation. Trees were generated separately for each essay score under each condition, which resulted in nine single regression trees. Figures 1–3 are for Rubric 1 score, Rubric 2 score, and total essay score, respectively. In each figure, three subfigures represent the three conditions. In general, even though a maximum of 15 layers could potentially be plotted, all the single trees ranged from five to seven levels, indicating that after five to seven levels, the increase in R^2 was smaller than .01 for all models. It is also worth noting that, even though all the top predictors were used to produce the single regression trees, not all would show up as the most differentiating features in a tree and some most differentiating features could appear at different layers.

Take Condition A, Rubric 1 as an example. The top level was "typo correction rate." This was consistent with the results in Table 9, in which this feature was shown to have the highest importance level. Typo correction rate, at its level of 4.6, split students' writing performance into two groups: a higher scoring group (mean Rubric 1 score = 2.5, $n = 102$) and a lower scoring group (mean Rubric 1 score = 1.9, $n = 114$). On the next level was "uncorrected spelling errors," which further

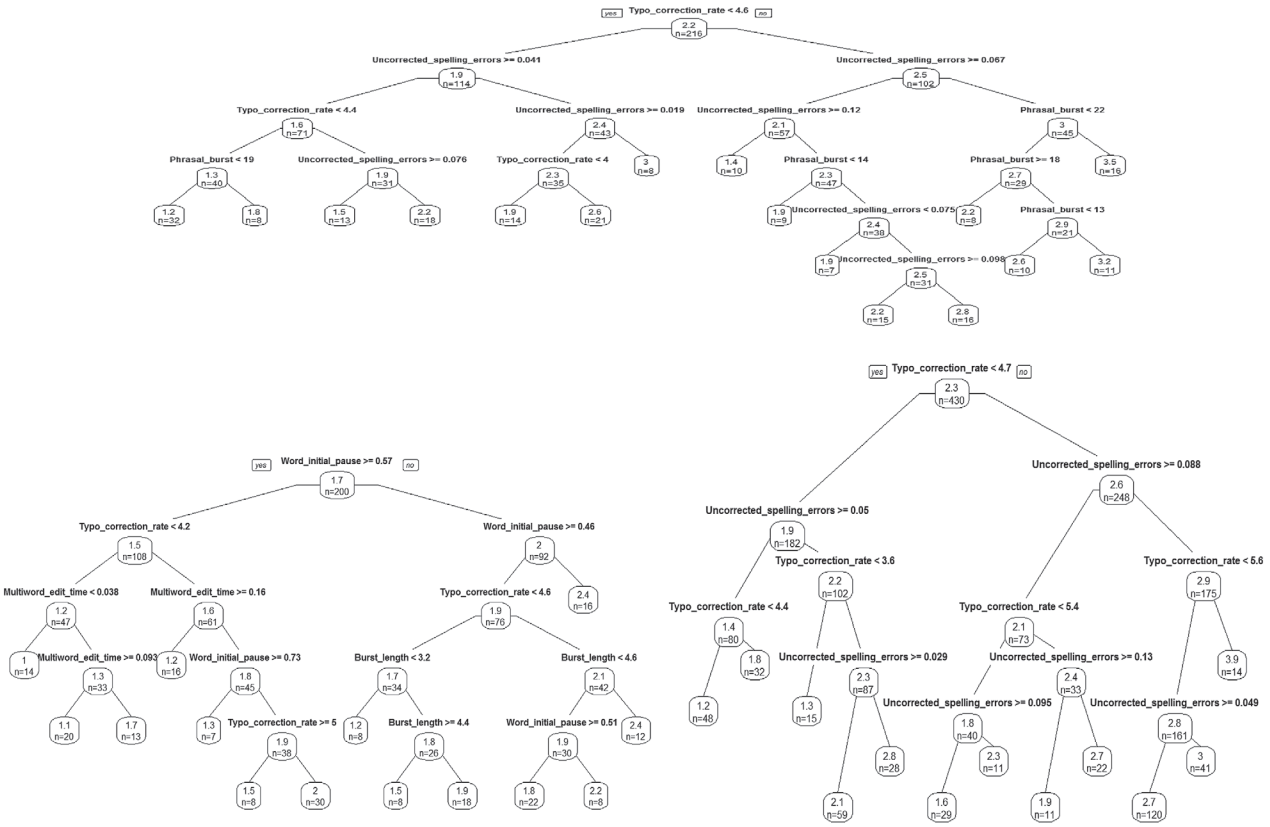

Figure 1 Regression tree plots for Rubric 1 score on three conditions: (top) Condition A, (bottom left) Condition B, and (bottom right) Condition C.

split the branches. Specifically, given that students had a typo correction rate < 4.6 (i.e., left branch), students with more uncorrected spelling errors (≥ .041) got lower mean Rubric 1 scores (i.e., 1.6, $n = 71$) than those with fewer uncorrected spelling errors (<.041; i.e., 2.4, $n = 43$). Note that moving down the tree level, each performance group started to include a smaller number of students. To continue, among students with a typo correction rate of ≥4.6 (i.e., right branch), students with more uncorrected spelling errors (≥.067) got lower essay scores on Rubric 1 (i.e., 2.1, $n = 57$), than students with fewer uncorrected spelling errors (<.067; i.e., 3, $n = 45$). The split of the tree continued through the list of predictors. As pointed out earlier, the same feature can appear at different levels. For example, "phrasal burst" appeared at Levels 3, 4, and 5. It is also possible that certain features included in Table 9 did not show up in the tree representations. For example, "characters in multiword deletion" was not shown in the tree, even though it was included in the model. This result suggested that this feature was not as effective or differentiating in splitting performance levels as other features, even though it was predictive of essay scores.

Note that the layers at lower levels of the regression trees should be interpreted with caution for at least two reasons. One reason is that the subgroup sample sizes at the lower branches tend to be rather small, hence the results may be affected largely by sampling error. The second reason is that the trees at the bottom layers are more likely to be false positive. As a result, with a different random sample, the identified variables and subgroups at the upper tree levels would tend to stay the same, but the variables at the lower tree levels might vary.

In general, across essay scores and assessment conditions, the results of the single trees were largely consistent with patterns reported earlier in Table 9. The features with the greater importance or relative influence measures appeared on the upper levels in the trees, which was another manifestation that they were more effective in differentiating writing performance. In particular, typo correction rate and uncorrected spelling errors were among the most important features in differentiating writing performance across conditions. These two features played a more dominant role as a splitting variable in Conditions A and C. Similar to the earlier reported results, Condition B was unique in that the feature

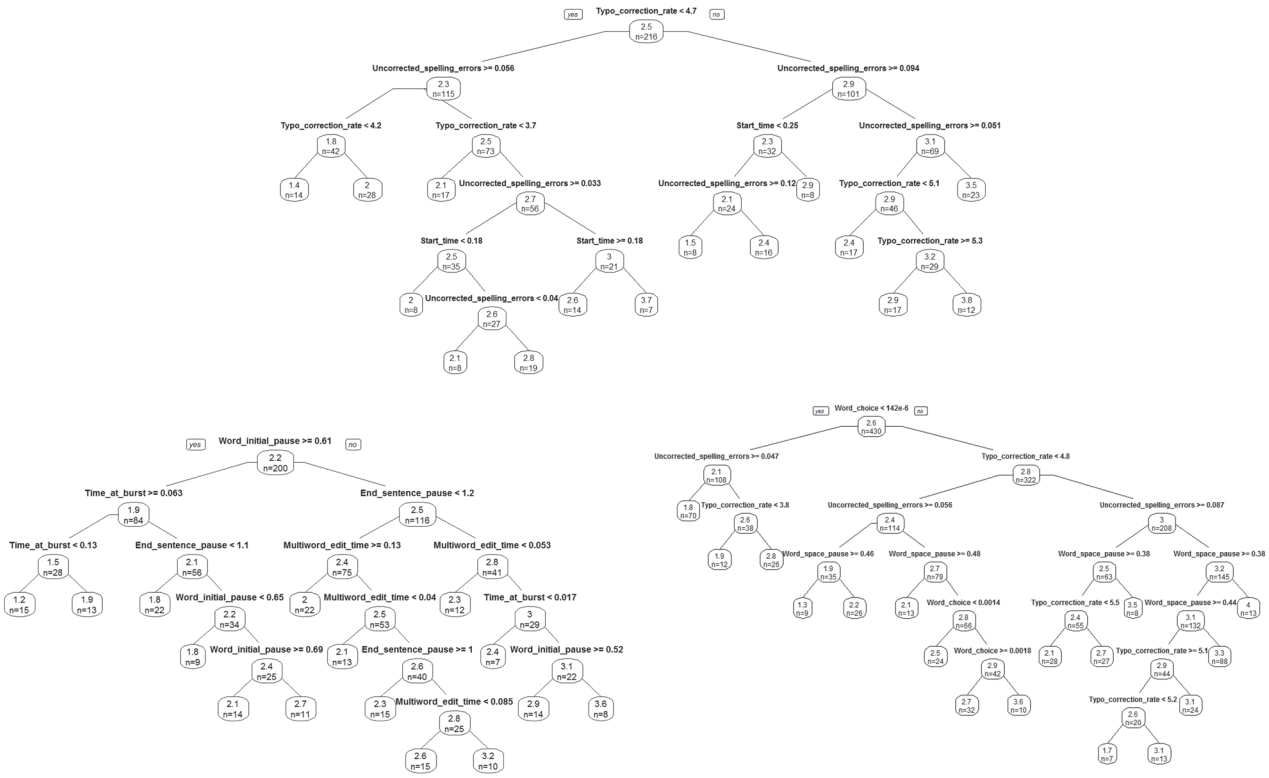

Figure 2 Regression tree plots for Rubric 2 score on three conditions: (top) Condition A, (bottom left) Condition B, and (bottom right) Condition C.

distinguishing writing performance the best was "word-initial pause," which did not manifest as the most effective splitting variable in other conditions.

Discussion

In this exploratory study, we addressed two research questions. In the first research question, we examined how essay scores differed when essay tasks were given in different assessment conditions. Three assessment conditions were examined: a base condition with scaffolding and topic effects, a condition with only the scaffolding effect, and a condition with neither effect. The results revealed significantly lower essay scores when the unified topic effect was removed compared to the original/base form condition and no significant essay score increase or drop when neither scaffolding nor topic effects existed, compared to the original form condition. This result is highly consistent with previous studies, such as Zhang, van Rijn, et al. (2019), and indicates that shifting the topic on an essay task to one that is different from the topic of the lead-in tasks can have a negative impact on students' writing performance in terms of getting lower essay scores. Students' essay scores were not statistically higher when they wrote the essay with the scaffolding support from a sequence of lead-in tasks on the same topic compared to when they were presented the essay task without much scaffolding and topic support.

However, several previous studies suggested that, regardless of essay score differences, students' writing processes could differ by assessment conditions. Therefore, in the second research question, we used tree-based methods to investigate which writing process features were more important, or had greater influence, in predicting students' writing performance under each of the three different assessment conditions. The writing process feature, typo correction rate, was consistently the highest ranking feature predicting the essay scores well, regardless of whether the essay task was given with the scaffolding and topic support. This result suggested that the rate of typo correction in the writing process had similar association to final essay quality regardless of the presence of scaffolding and/or topic effects. Moreover, Condition B, in which the culminating essay task was on a different topic from the lead-in tasks, behaved differently from other conditions. Specifically, one feature, uncorrected spelling error, ranked high for Conditions A and C, but not for Condition B, in which the topic effect was broken. Condition B, instead, called on the importance of word-initial pause and multiword

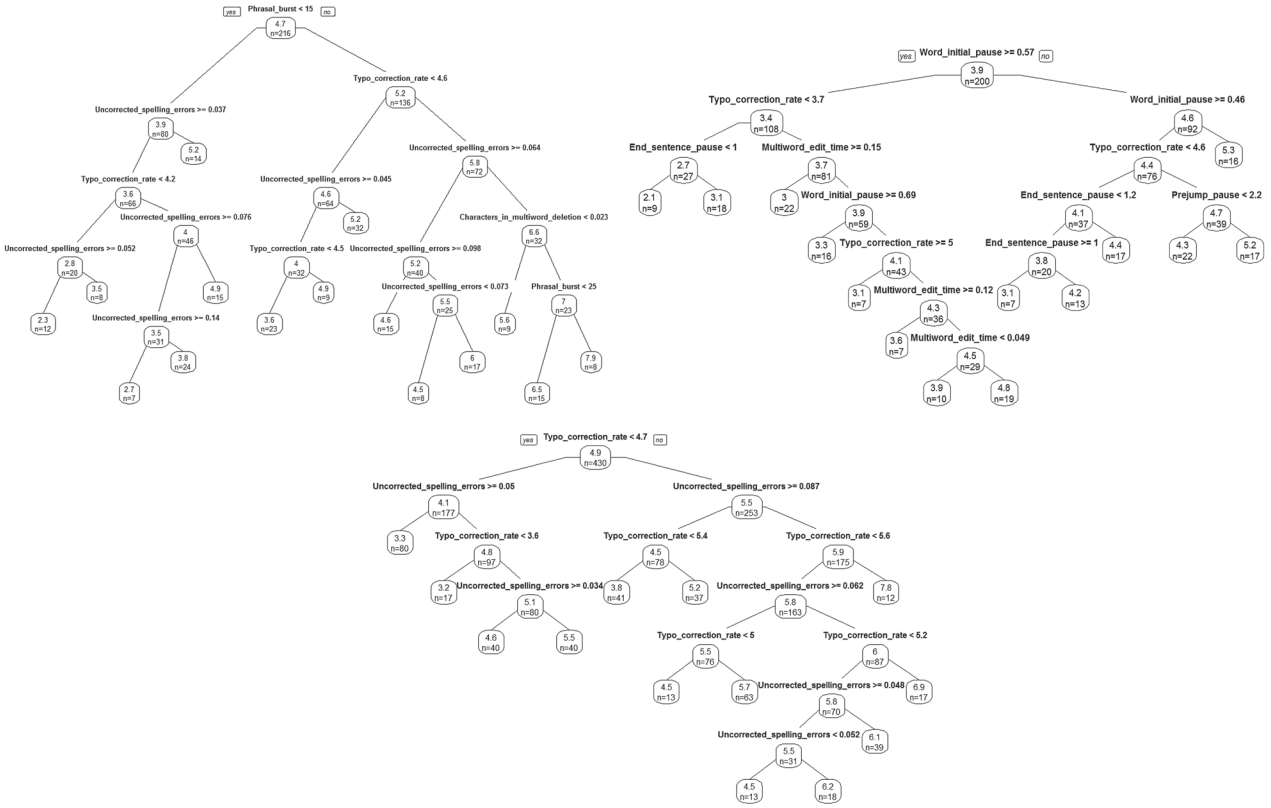

Figure 3 Regression tree plots for total essay score on three conditions: (top left) Condition A, (top right) Condition B, and (bottom) Condition C.

edit time, which was largely absent for the other two conditions. We hypothesize that the sudden shift of topic in the essay task had caused cognitive burden to the students, leading them to place more emphasis on word-level finding and editing, even though the same source materials were supplied in Condition B prior to essay writing. Additionally, the quality of the final response tended to rely more on phrase and sentence-level editing and planning when the topic effect was broken in an assessment, possibly due to the lack of familiarity with the topic further exacerbated by the shift in topic. All in all, the results of this study show that different assessment conditions (test designs) call on different configurations of writing processes. The most salient and noteworthy result is that the broken topic effect might affect the extent of word finding and retrieval as well as editing across multiple words by the students.

This study provides additional evidence on the impacts of scenario-based assessment design on students' essay writing processes. Following previous studies that reported students writing more efficiently under scenario-based assessments (e.g., Zhang, van Rijn, et al., 2019), this study further unpacks the writing processes under such assessment design using features extracted from keystroke logs. While the study answers some research questions, it also motivates new study directions. Given the findings, one direction of future study is to conduct an in-depth analysis on the vocabularies used by students in different conditions and see how relevant and specific the vocabularies in different conditions are to the writing topic — and the extent of misspelling rate on topic-specific vocabularies. Another, more straightforward approach is to conduct cognitive interviews with students who can then verbalize how the support from the lead-in tasks might have affected their idea generation and text production processes. Additionally, it is unclear how the topic and task ordering effects are the same for students with different demographic backgrounds or for students of different ability levels. The theoretically ideal condition — Condition A — may be more effective with students of lower ability levels than with students of high ability levels. It would be worthwhile to study the form effects on different student populations.

Despite the interesting findings and the use of new analytic methods, this study has several limitations. First, voluntary student samples in a single grade level were used for analyses. The data were collected under a low-stakes testing setting, which might have affected students' motivation and performance. Second, single regression trees generally suffer from high variance and low predictive accuracy, and that is why boosted regression trees were first used to select key predictors.

In this study, we used an incremental R^2 of .01 as a predictor selection criterion, which may be considered generous. As a result, the lower levels of the trees may not be as stable as high levels of the trees; therefore, our results should be interpreted with caution.

Acknowledgments

We thank Paul Deane and Randy Bennett for providing technical guidance on the study design and analyses. We also thank the editors and reviewers for their insightful suggestions on the previous version of this manuscript. Jianshen Chen is now working at the College Board.

Note

1 Van Rijn and Yan-Koo (2016) analyzed these parallel forms.

References

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, *8*(2–3), 70–91. https://doi.org/10.1080/15366367.2010.508686

Bennett, R. E., Deane, P., & van Rijn, P. W. (2016). From cognitive domain theory to assessment practice. *Educational Psychologist*, *51*(1), 82–107. https://doi.org/10.1080/00461520.2016.1141683

Cayton-Hodges, G. A., Marquez, E., Keehner, M., Laitusis, C., van Rijn, P., Zapata-Rivera, D., Bauer, M. I., & Hakkinen, M. T. (2012). *Technology enhanced assessments in mathematics and beyond: Strengths, challenges, and future directions*. Educational Testing Service.

Chen, J., & Keller, B. (2019). Heterogeneous subgroup identification in observational studies. *Journal of Research on Educational Effectiveness*, *12*(3), 578–596. https://doi.org/10.1080/19345747.2019.1615159

Deane, P. (2014). *Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks* (Research Report No. RR-14-03). Educational Testing Service. https://doi.org/10.1002/ets2.12002

Deane, P., Feng, G., Zhang, M., Hao, J., Bergner, Y., Flor, M., Wagner, M., & Lederer, N. (2016). *Generating scores and feedback for writing assessment and instruction using electronic process logs* (U.S. Patent No. 14/937,164). U.S. Patent and Trademark Office.

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). Educational Testing Service.

Deane, P., Sabatini, J. P., Feng, G., Sparks, J., Song, Y., Fowles, M., O'Reilly, T., Jueds, K., Krovetz, R., & Foley, C. (2015). *Key practices in the English language arts (ELA): Linking learning theory, assessment, and instruction* (Research Report No. RR-15-17). Educational Testing Service. https://doi.org/10.1002/ets2.12063

Deane, P., & Song, Y. (2015). *The key practice, "Discuss and debate ideas": Conceptual framework, literature review, and provisional learning progressions for argumentation* (Research Report No. RR-15-33). Educational Testing Service. https://doi.org/10.1002/ets2.12079

Deane, P., Song, Y., van Rijn, P., O'Reilly, T., Fowles, M., Bennett, R., Satatini, J., & Zhang, M. (2018). Scenario-based assessment of argumentation. *Reading and Writing*, *32*, 1575–1606. https://doi.org/10.1007/s11145-018-9852-7

Deane, P., & Zhang, M. (2015). *Exploring the feasibility of using writing process features to assess text production skills* (Research Report No. RR-15-26). Educational Testing Service. https://doi.org/10.1002/ets2.12071

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, *22*(9), 1365–1381. https://doi.org/10.1002/sim.1501

Greenwell, B., Boehmke, B., Cunningham, J., & Metcalfe, P. (2019). *gbm: Generalized boosted regression models* [Computer software manual]. https://cran.r-project.org/package=gbm

Guo, H., Zhang, M., Deane, P., & Bennett, R. (2020). *Effects of scenario-based assessment on students' writing processes. Journal of Educational Data Mining*, *12*(1), 19–45. https://doi.org/10.5281/zenodo.3911797

Inputlog [Computer software]. (2013). http://www.inputlog.net/

jEdit [Computer software]. (1998). http://www.jedit.org/

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, *30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Li, C., Zhang, M., & Deane, P. (2016, April 11). *Investigating the relations of writing process features and the final product*. Paper presented at the National Council on Measurement in Education, Washington, DC.

O'Reilly, T., & Sabatini, J. P. (2013). *Reading for understanding: How performance moderators and scenarios impact assessment design* (Research Report No. RR-13-31). Educational Testing Service. https://doi.org/10.1002/sim.1501

Quellmalz, E. S., Davenport, J. L., Timms, M. J., DeBoer, G. E., Jordan, K. A., Huang, C.-W., & Buckley, B. C. (2013). Next-generation environments for assessing and promoting complex science learning. *Journal of Educational Psychology*, *105*(4), 1100–1114. https://doi.org/10.1037/a0032220

R Core Team. (2015). *R: A language and environment for statistical computing* [Computer software manual]. https://www.r-project.org/

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, *31*, 172–181.

ScriptLog [Computer software]. (2006). http://www.scriptlog.net/

Sinharay, S., Zhang, M., & Deane, P. (2019). Prediction of essay scores from writing process and product features using data mining methods. *Applied Measurement in Education*, *32*(2), 116–137. https://doi.org/10.1080/08957347.2019.1577245

Sullivan, K. P. H., & Lindgren, E. (Eds.) (2006). *Computer key-stroke logging and writing: Methods and applications*. Elsevier.

Therneau, T., Atkinson, B., & Ripley, B. (2019). *Rpart: Recursive partitioning and regression trees* [Computer software manual]. https://cran.r-project.org/web/packages/rpart

Translog [Computer software]. (2006). http://www.translog.dk/

Van Rijn, P., & Yan-Koo, Y. (2016). *Statistical results from the 2013 CBAL English language arts multistate study: Parallel forms for argumentative writing* (Research Memorandum No. RM-16-15). Educational Testing Service.

Zhang, M., Bennett, R., Deane, P., & van Rijn, P. (2019). Are there gender differences in how students write their essays? An analysis of writing processes. *Educational Measurement: Issues and Practice*, *38*(2), 14–26. https://doi.org/10.1111/emip.12249

Zhang, M., & Deane, P. (2015). *Process features in writing: Internal structure and incremental value over product features* (Research Report No. RR-15-27). Educational Testing Service. https://doi.org/10.1002/ets2.12075

Zhang, M., van Rijn, P. W., Deane, P., & Bennett, R. E. (2019). Scenario-based assessments in writing: An experimental study. *Educational Assessment*, *24*(2), 73–90. https://doi.org/10.1080/10627197.2018.1557515

Zhang, M., Zou, D., Wu, A., Deane, P., & Li, C. (2017). An investigation of writing processes employed in scenario-based assessment. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 321–339). Springer. https://doi.org/10.1007/978-3-319-56129-5_17

Suggested citation:

Cao, Y., Chen, J., Zhang, M., & Li, C. (2020). *Examining the writing processes in scenario-based assessment using regression trees* (Research Report No. RR-20-18). Educational Testing Service. https://doi.org/10.1002/ets2.12301

Action Editor: James Carlson

Reviewers: Sandip Sinharay and Mengxiao Zhu