


The case for scenario-based assessment of written argumentation

Paul Deane¹  · Yi Song¹ · Peter van Rijn¹ · Tenaha O'Reilly¹ · Mary Fowles¹ · Randy Bennett¹ · John Sabatini¹ · Mo Zhang¹

Published online: 7 July 2018

© Springer Science+Business Media B.V., part of Springer Nature 2018

Abstract This paper presents a theoretical and empirical case for the value of scenario-based assessment (SBA) in the measurement of students' written argumentation skills. First, we frame the problem in terms of creating a reasonably efficient method of evaluating written argumentation skills, including for students at relatively low levels of competency. We next present a proposed solution in the form of an SBA and lay out the design for such an assessment. We then describe the results of prior research done within our group using this design. Fourth, we present the results of two new analyses of prior data that extend our previous results. These analyses concern whether the test items behave in ways consistent with the learning progressions underlying the design, how items measuring reading and writing component skills relate to essay performance, how measures of transcription fluency and proficiency in oral and academic language relate to writing skill, and whether the scenario-based design affects the fluency and vocabulary used in an essay. Results suggest that students can be differentiated by learning progression level, with variance in writing scores accounted for by a combination of performance on earlier tasks in the scenario and automated linguistic features measuring general literacy skills. The SBA structure, with preliminary tasks leading up to the final written performance, appears to result in more fluent (and also more efficient) writing behavior, compared to students' performances when they write an essay in isolation.

Keywords Assessment · Argumentation · Scenario-based assessment · SBA · Writing · Summary · Summarization · Essay · Critique · Reading · IRT · Dimensionality · Learning progression · Keystroke · Burst · Process data · Automated essay scoring · AES

✉ Paul Deane
pdeane@ets.org

¹ Educational Testing Service, Rosedale Road, MS 11-R, Princeton, NJ 08540, USA

Introduction

The problem

The ability to engage in thoughtful, constructive argument is a critical goal in twenty-first-century education (Goldman et al., 2016), as reflected by its importance in modern educational standards. For instance, the Common Core State Standards, or CCSS (CCSSO & NGA, 2010) have set ambitious goals for reading and writing arguments (CCSS Writing Strand 3). Yet the evidence suggests that students rarely produce effective written arguments (Kuhn, 1991; Shemwell & Furtak, 2010).

Of course, argument is a complex performance skill that requires coordination of various component skills (Kuhn, 1991), which can be difficult to teach and challenging to assess (Kuhn & Crowell, 2011; Mayweg-Paus, Macagno, & Kuhn, 2016). The best assessments of argument skill require students not only to construct arguments about specific issues, but also to consider alternate perspectives and evaluate the complex arguments of others (Kuhn & Udell, 2007). Although it is often difficult to detect the causes of weak arguments, assessment, ideally, should help teachers identify where students encounter difficulties with argument and set specific targets for improvement.

In accountability contexts, assessments of written argument have historically fallen into three categories: (1) selected-response assessments, (2) constructed-response assessments, and (3) projects and portfolios.

Selected-response tests pose specific problems and ask students to choose the best solution(s) from a list of options. Items cover a range of skills, extending from writing fundamentals (e.g., grammar, syntax, and sentence structure) to argument quality. These assessments can gather evidence about a variety of elemental skills relatively quickly, but they do not exercise critical aspects of written argument or other advanced literacy skills, specifically the ability to integrate and manage the multiple competencies necessary to produce effective arguments.

Constructed-response assessments that require students to respond to a short prompt in a fixed time, as in the National Assessment of Educational Progress (NAEP) and other standardized tests, support the evaluation of integrated, holistic performance. In typical implementations, though, students have little time to develop their understanding of the topic, and only the performance product is measured, not the processes and components that contribute to it. For these reasons, among others, timed constructed-response assessments have frequently been criticized (Hillocks, 2002; White, 1995). Such measures often provide a single score, but when they report more detailed results, e.g., subscores, these tend to be highly correlated (Godshalk, Swineford, & Coffman, 1966; Levy, 2013) and therefore provide little unique information that can differentially guide instruction. This concern is particularly relevant for students who write significantly less than their more skilled counterparts (Ferrari, Bouffard, & Rainville, 1998). It may be hard to determine what a low score on an argument performance assessment implies about students' ability to generate arguments, analyze positions, claims, or supporting evidence, or comprehend source texts.

Finally, projects and portfolios closely resemble (or directly record) work done in school or professional settings. Sources must be read, the arguments of others summarized, analyzed, and critiqued; a position developed, and one's argument expressed using evidence from sources. The challenge from an accountability assessment perspective is in the cost, time, logistics, and complexity of interpreting such measures.

The proposed solution

This paper presents a class of *scenario-based assessments* (SBAs) designed to support and measure written argument, an approach that addresses several of the limitations discussed previously. We review past research on SBAs and present two new analyses that address the quality of these measures.

In an SBA, students are given a purpose for reading a collection of thematically related texts. Tasks, activities, and resources are sequenced to develop students' understanding of an issue as they engage in increasingly more complex reasoning tasks. As part of this process, an SBA measures the component skills that feed into the culminating task (e.g., a writing task that requires students to integrate and demonstrate knowledge about the issue under discussion). The results identify specific parts of the larger task students can or cannot do, thus providing teachers with information they can use to build on student strengths and to address their weaknesses.

SBAs embody some of the best features of the three assessment categories described above. As we conceive them, SBAs have the following features: first, the task sequence simulates a condensed writing project undertaken in an order that a skilled practitioner might follow. In effect, the SBA models what we term a *key practice*, the coordinated execution of a bundle of skills commonly exercised in writing communities for which students are being educationally prepared (Deane et al., 2015). Second, each task assesses a skill that contributes to success on the practice as a whole; and third, each individual item within a task is designed to measure whether students have reached a specific level in one or more *learning progressions*. A learning progression is based upon a theory of the domain and is designed to provide meaningful descriptions of student progress toward mastery of an important aspect of the key practice (see Deane & Song, 2014, 2015).

SBAs have been shown to provide valid indicators of reading comprehension across a range of ability levels for elementary (Sabatini, Halderman, O'Reilly, & Weeks, 2016), middle (Sabatini, O'Reilly, Halderman, & Bruce, 2014), and high school students (O'Reilly, Weeks, Sabatini, Halderman, & Steinberg, 2014). The cited studies indicate that SBAs can measure both complex reading processes, such as the ability to integrate and evaluate multiple sources in a digital environment, and component processes, including mental-model formation, source evaluation, perspective taking (Sabatini et al., 2014), and background knowledge (O'Reilly et al., 2014; Sabatini et al., 2016) while providing valid measurement.

With respect to written argument, our concern is how well-prepared students are to carry out a literacy task that requires them to first read and evaluate arguments, then develop and present their own written position on the issue. To support this

goal in an effective assessment, we first analyzed the domain to identify critical tasks and component skills. Next, we designed scenarios to guide students through a task sequence enacting the most important of these competencies. Finally, we created test forms (hereafter referred to as *forms*) that embodied the resulting design. Key features of our domain analysis, task selection, and assessment designs are shown below.

Assumptions about written argument

Various empirical studies have examined the development of argument skills across school years. The development of argument depends on the emergence of critical subskills, such as selecting relevant evidence (Brem, 2000; Kuhn, 1991; Kuhn, Shaw, & Felton, 1997). McCann (1989) found that young children can express their opinions and offer supporting reasons. According to Ferretti et al. (2009), upper-elementary school students can elaborate and provide details to support their arguments in writing. Similarly, Kuhn and Crowell (2011) found that training can help sixth graders become more aware of the need to use relevant evidence to support their claims. However, some argument skills are challenging, even in the upper grades, and may not develop before adulthood unless support is provided. For example, Kuhn (1991) and Klaczynski (2000) found that high school and college students have difficulty identifying the assumptions behind people's arguments and to integrate arguments from various sides of an issue. Even adults have trouble refuting opposing viewpoints and anticipating counterarguments, especially in a written context (Leitão, 2003; Nussbaum & Kardash, 2005).

The assessment design discussed in this paper presupposes these developmental patterns and focuses on two critical skills: the ability to create and evaluate arguments and to summarize arguments in informational texts. Specifically, the design is based on a theoretical framework with learning progressions proposed for argument (Bennett et al., 2016; Deane & Song, 2014) and for informational reading and writing, including summary (O'Reilly, Deane, & Sabatini, 2015). We hypothesized that achievement in written argument is closely linked to student progress in both skillsets and built SBAs to test this hypothesis at the middle-school level.

Specifics of the assessment design: how learning progressions and task sequences can help deconstruct student performance

The assessments we describe are comprised of items intended to measure student performance on learning progressions (LPs) for summary and argument within a unifying scenario.

Summary learning progressions

The summary LP has five levels. At Level 1, students have some ability to recognize what information is present and salient in a text, but relatively little grasp of text structure. At Level 2, students can represent text structure to some degree, but may

have difficulty expressing both main and subordinate points clearly when they try to summarize a text. At Level 3, students can recognize and describe text structure in a summary but may have difficulty inferring textually implicit information or evaluating their summary for accuracy of information and originality of expression. At Level 4, students demonstrate stronger inference and evaluation skills, though they may take relatively rigid approaches to summary, without adapting their summary strategies to the disciplinary context, the skill needed to reach Level 5. More details on the theoretical framework and specific descriptors of our summary learning progression are provided in O'Reilly et al. (2015). The assessments we describe include summary items that target Levels 1–4, as summarized in Table 1.

Argument learning progressions

The argument LP also has five levels. At Level 1, students can contribute single turns to an ongoing oral argument, such as making a claim, providing supporting evidence, or raising an objection, but their argument skills may be entirely tacit, without metacognitive awareness of argument structure. At Level 2, students have more metacognitive control over argument, sufficient to build a simple case consisting of claims and reasons, but may not make effective use of evidence to evaluate or strengthen arguments. At Level 3, students can construct and present a multi-level argument coordinating claims, reasons, and evidence, but they may still exhibit my-side biases or fail to provide objective critiques, which are Level 4 descriptors. At Level 5, most characteristic of college or, more likely, postgraduate levels of performance, students can handle the complexities of many-sided argumentative discourses where one's own assumptions and presuppositions may constantly be challenged. Details on the theoretical framework underlying our Argument Learning Progression are provided in Deane and Song (2014, 2015). The assessments we describe target Levels 1–4 (see Table 2).

Table 1 Summary task descriptions

Task description	Learning progression level	Number of items
Distinguish opinion from statement of fact	1	1 Selected-response item
Distinguish detail from main idea	1	1–2 Selected-response items
Identify the main idea	1	1 Selected-response item
Identify supporting ideas	2	1 Selected-response item
Write a summary	3	2 Constructed-response items
Evaluate accuracy of information in summary	4	1–2 Selected-response items
Recognize plagiarism in summary	4	1–2 Selected-response items

Table 2 Argument task descriptions

Task description	Learning progression level	Number of items
Classify reasons as being for or against a position	1	1 Task composed of 10 related items
Recognize evidence supporting a claim	1	2–3 Selected-response items (out of 6)
Recognize evidence that weakens a claim or which neither supports nor weakens a claim	2	2–3 Selected-response items (out of 6)
Write an argumentative essay that includes a clear position, multiple supporting reasons, and some relevant evidence	3	1 Constructed-response item
Evaluate and critique an argument	4	1 Constructed-response item

Sequence of lead-in tasks

The form described in Tables 1 and 2 has four sections, intended to recapitulate steps an experienced writer might follow to prepare for and write an argument essay. In Section One, students read and summarize articles about an issue (a set of tasks designed not only to assess summary but to also build up content understanding to support the essay-writing task). In Sections Two and Three, students analyze arguments (e.g., writing a critique of a letter with flawed arguments and evaluating whether evidence supports or weakens an argument). In Section Four, they write an essay of their own.

The lead-in tasks were designed to isolate components that feed into the culminating essay-writing performance. In particular, they were intended to help disambiguate whether low essay scores reflect an inability to compose arguments in written form, an inability to analyze key aspects of an argument (e.g., detect logical errors), or difficulties in understanding and re-presenting information from source articles (e.g., the summary tasks).

This design was initially developed as part of a collaboration between the authors and a school district in the northeastern United States. Initial design and development work is described in Deane, Fowles, Baldwin, and Persky (2011) for a form focused on the topic, “Should the United States government ban advertising to children under twelve?” (hereafter referred to as “Ban Ads”). For subsequent studies, we created both parallel and adapted forms to address a variety of research questions. Below we report two studies that use data collected in previous investigations. The results of these and related investigations are summarized in the Instrument sections for each study.

Study 1

In Study 1 we examined whether individual items in the SBA design worked as intended so that patterns of performance could be presented to teachers to help them identify instructional priorities, especially for students who perform poorly on the essay. In particular, we addressed the following research questions:

1. *Are the observed difficulties of individual items consistent with their assigned learning progression levels in our theoretically driven assessment design?* This question focuses on the validity of prior LP item-level assignments and the sequence of item-level difficulty. To the extent that the LP levels can be validated, they may provide guidance about specific skills that lower-performing students most urgently need to master.
2. *Does each component reading and writing lead-in task contribute unique variance to the predicted student performance on the culminating essay-writing task?* If the lead-in tasks measure skills associated with written argument, particular tasks might have diagnostic value, especially when students complete them successfully but do not produce a well-reasoned essay.
3. *To what extent do transcription fluency and proficiency in oral and academic language relate to the quality of written argument?* This question examines the extent to which features that could be automatically extracted from student performance provide information about basic writing skills (which may also account for lower performance on written argument tasks).

Method

Since Study 1 includes additional analyses on data used in van Rijn and Yan-Koo (2016), we describe the instruments, population, and some of their selected results as background.

Participants

Data were collected in 2013 from 382 7th-grade students, 913 8th-grade students, and 537 9th-grade students from 18 schools in six states. The largest group (58%) was from one western U.S. state. Demographic data were available for 70% of students. Of this subset, 65.6% were White, 16.6% Hispanic, 3.4% Asian, and 3.2% African American. There were slightly more females (46.9%) than males (43.2%), with the remainder unreported. Less than 1.9% were reported as former or current ELLs. Finally, 19.2% were reported as receiving free or reduced-price lunch.

Instrument and procedure

Three parallel argument forms were used; each form replicated the design and scenario structure described above but focused on a different topic. These forms included the original Ban Ads form and two additional forms focusing on the

questions: “Should students be given cash rewards for getting good grades?” (Cash for Grades) and “Should schools encourage parents to place limits on students’ use of social networking sites?” (Social Networking).

Each form was administered during two online, 45-min sessions. The summary and critique tasks were administered in the first session, and the argument analysis and essay tasks in the second. The two sessions were administered as closely together as feasible, with no more than 1 week intervening. Students were assigned randomly within classrooms to one of six counter-balanced administration sequences of form pairs.

Total scores had reasonable levels of reliability, with internal consistency coefficients (Cronbach’s alpha) between .81 and .83, and correlations between pairs of parallel forms between .73 and .80.

Rater agreement on constructed-response items on each form (Fu, Chung, & Wise, 2013; Fu & Wise, 2012; van Rijn & Yan-Koo, 2016) had quadratic weighted kappa values between .68 and .76 for the summary items, .87 to .89 for the critique items, .77 to .84 for the first essay rubric (a 5-point scale for general writing qualities such as organization, development, vocabulary, grammar, usage, mechanics, and style), and .80 to .83 for the second essay rubric (a 5-point scale focused on the quality of argument) (van Rijn & Yan-Koo, 2016).

With respect to test design, results from dimensionality analyses were consistent with the hypothesized constructs. Several models were fit to the items from the three forms. A multidimensional item-response theory model had the best fit with distinguishable, but strongly correlated dimensions ($r = .76$) for the selected-response items (measuring reading skills) and constructed-response (writing) items. A second model that postulated dimensions by scenario topic did not fit as well and yielded almost perfectly correlated dimensions ($r > .97$). These results suggested that, as intended, items created to measure argument reading and argument writing skills represented related, but separable sub-constructs. Also, the issue that students were asked to address did not have a major impact on measurement properties, though item difficulty and discrimination varied somewhat across forms.

In another study, van Rijn, Graf, and Deane (2014) used overall test results to assign students to Argument Learning Progression (LP) levels on each of the two forms a student took, employing “task progression maps” to link segments of the underlying ability scale with LP levels, and assigning the cut score for each level as the point at which 65% of the students performed at expected levels on each task. Results showed reasonable classification agreement. For each pair of forms, about half the students were assigned to the same LP level, and more than 90% were assigned to the same or adjacent LP levels. In this model, the items only measured Levels 1–4; some students performed below the cut score for Level 1 and were thus were classified as falling below Level 1.

Finally, reasonable relations with other indicators of reading and writing skill were found, with observed correlations with state reading test scores between .58 and .61, and with state language-use test scores between .56 and .60. In another study, Zhang, Zou, Wu, Deane, and Li (2017) found a correlation of .53 between the Ban Ads form total score and teacher ratings of student writing ability. Additional

results relating to these forms can be found in Bennett (2011), Fu et al. (2013), and Fu and Wise (2012).

Data analysis

To answer Research Question 1, focusing on whether empirical item statistics were consistent with the LP-based design, we conducted an item-level (and in some cases, option-level) analysis for all items, disaggregating by student LP level. For selected-response questions with a binary correct/incorrect score, we calculated percent correct by student LP level. For selected-response items with multiple answers, we calculated the percentage of students who correctly selected each option. For constructed-response items, we calculated the percentage of students who scored at or above the value required to demonstrate the item's assigned LP level. Following the operational definition of a cutoff for LP levels employed by van Rijn et al. (2014), we hypothesized that at least 65% of students *at* the LP level assigned to an item would answer it correctly—and that most students *below* that LP level would not. The basis for using a cutoff at 65% lay in a standard-setting method in which segments of item-response probabilities ranging from .50 to .80 are placed on the ability scale using an IRT model (Van der Schoot, 2002). The 65% cutoff represented the midpoint of these segments.

To answer Research Question 2, focusing on whether the component reading and writing tasks predict essay performance, we conducted multiple linear regression analysis, regressing essay performance on the other task scores. We hypothesized that each major task (summary reading, summary writing, argument analysis, and critique writing) would contribute unique variance to the prediction of essay score.

To answer Research Question 3, focusing on the relationship of transcription fluency and language proficiency to the quality of written argument, we examined the extent to which fluency of typing and linguistic quality indicators predicted essay score.

The literature indicates that *bursts* (sequences of fast typing without any intervening long pauses) provide evidence about the fluency of idea generation and sentence-planning processes. Greater writing fluency corresponds to longer, more variable bursts of text production. Less fluent writing tends to contain a larger proportion of very short bursts, which might be the result of weaker transcription skills or of working memory loads imposed by competing cognitive processes (Alves & Limpo, 2015; Hayes, 2012). We analyzed the digitally captured keystroke logs of all students to extract the following measures: (a) the number of bursts of text production in each log, using a cut-off of two-thirds of a second to define burst boundaries, as specified by Almond, Deane, Quinlan, and Wagner (2012); (b) the mean log length of bursts in characters; (c) burst pacing (the extent to which writers produce long bursts quickly), defined as the sum of log burst length normalized against total time on task. Details of the methods used to capture these features are documented in Almond et al. (2012).

To measure qualitative differences in the linguistic properties of student essays, we used feature scores derived from the e-rater[®] automated scoring engine, which have been shown to predict student essay scores accurately (Attali & Burstein,

2005). These scores enabled us to capture linguistic correlates of writing quality, including avoidance of grammar, usage, and mechanics errors; avoidance of stylistic faults (e.g., repetition of words); idiomaticity of language, as measured by the presence of common collocations and the avoidance of preposition errors; difficulty of vocabulary, as measured by the lower frequency and increased word length; and syntactic variety, as measured by the rate at which a range of grammatical categories were produced.

We hypothesized that ability in written argument would be predicted by greater writing fluency and the presence of these linguistic features, since transcription skill and verbal ability are critical foundational skills for written argument. To assess this hypothesis, we conducted latent variable regressions, estimating student ability in a unidimensional IRT model but entering these measures of fluency and linguistic sophistication as predictors.

Results

Research Question 1: empirical fit of argument items to the argument learning progression

Our assessment design mapped onto four argument LP levels: Level 1 (the pro/con argument task), Level 2 (the strengthen/weaken-argument task), Level 3 (the essay-writing task), and Level 4 (the critique-writing task). We therefore expected to classify students into five groups: students below Level 1 and at each of the four levels.

Level 1

The pro/con argument task required ten binary judgments. Our modelling framework postulated that students at LP Level 1 should achieve around 65% accuracy on the task as a whole, which implied that they should reach that level on well over half the individual binary choices. As indicated in Fig. 1, which shows the percentage-correct for all 30 individual statements that students must classify across the three forms, students achieved the 65% percent-correct threshold for 27 of the 30 options at Level 1, but for less than half of the items below Level 1. The forms differed in difficulty for students below Level 1: Social Networking was harder than the other forms and Cash for Grades was easier, but this difference diminished at higher LP levels.

Level 2

The remaining argument analysis tasks addressed LP Level 2 and assessed whether students consistently recognized when evidence strengthens, weakens, or is irrelevant to an argument. However, our framework implied that items correctly answered by selecting the option “strengthen” should be easier than items correctly answered by the options “weakens” or “neither” (meaning irrelevant), since strengthen items did not require consideration of alternate perspectives. Our results

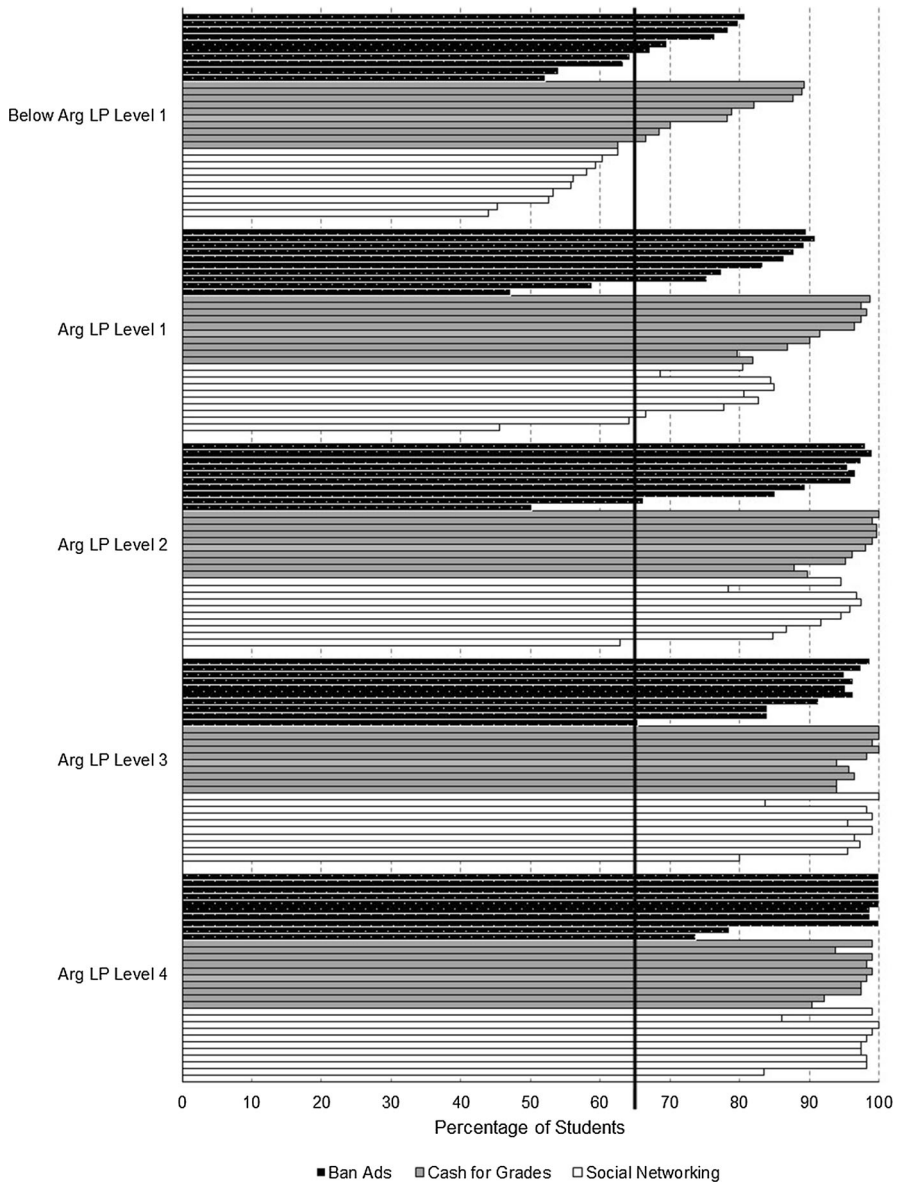


Fig. 1 Percentage of students who classified arguments accurately pro versus con by argumentation LP level. Each band represents judgments pro/con on a single statement. 65% threshold is indicated by black line

supported this prediction. As Fig. 2 illustrates, seven of eight “strengthen” items reached the 65% threshold at LP Level 1, though one was unexpectedly difficult. On the other hand, “weaken” and “irrelevant” items were harder. Only five of these ten items achieved 65% threshold at Level 1, whereas eight of the ten achieved the 65%

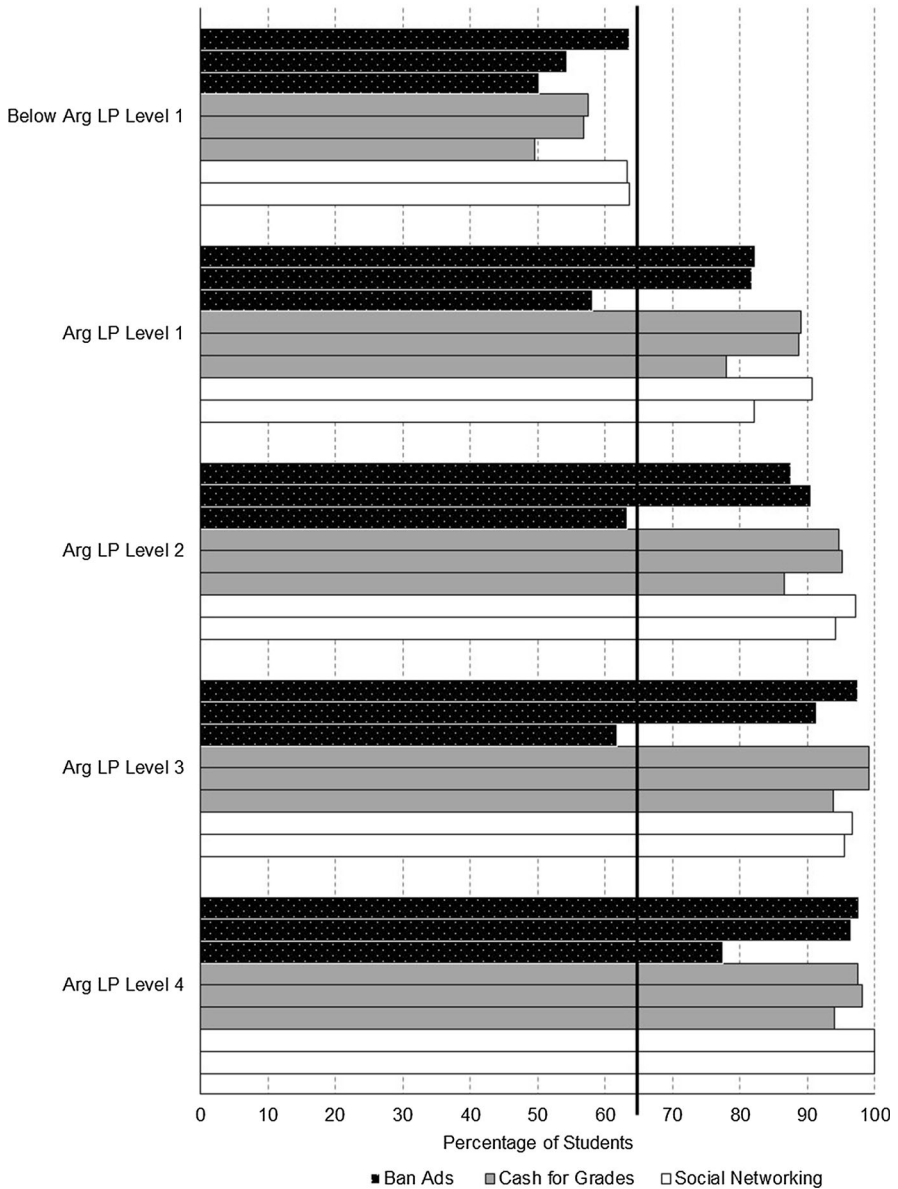


Fig. 2 Percentage of students who correctly identified strengthening evidence by argument LP level. The black line indicates the 65% threshold. *Arg LP* Argument learning progression

threshold at LP Level 2 (see Fig. 3). (Since the early assessment design did not specify parallel keys—i.e., correct answers, the number of strengthen, weaken, and irrelevant keys varied somewhat across forms.)

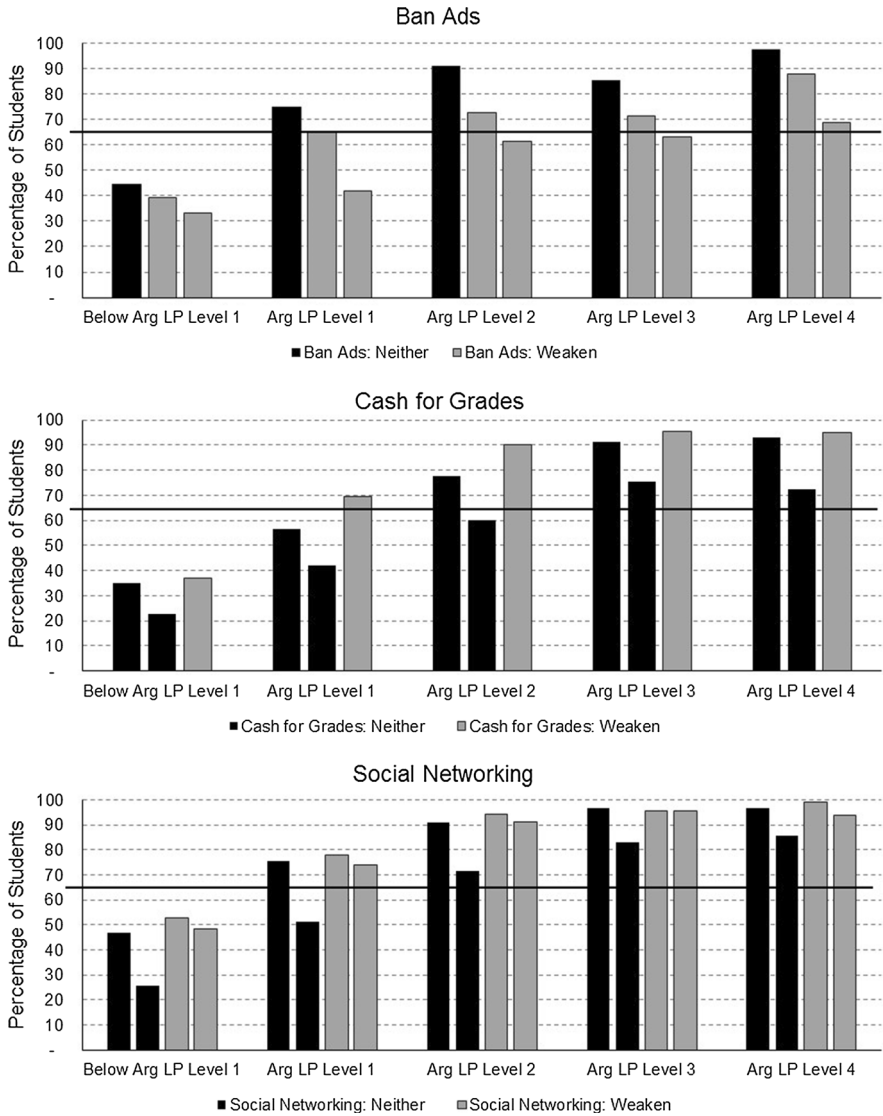


Fig. 3 Percentage of students who correctly identified weakening and irrelevant evidence by LP level. *Arg LP* Argument learning progression. The black line indicates the 65% threshold

Levels 3 and 4

The essay and critique writing tasks were even more difficult. We expected 65% of students at LP Level 3 to achieve at least seven of ten points available on the essay task, and 65% of students at LP Level 4 to consistently achieve at least three of the four points available on the critique task. Since these tasks were used to define LP Levels 3 and 4 in van Rijn et al. (2014), it is unsurprising that two of the three essay

tasks reached the 65% level at LP Level 3 as expected, with the third task falling only slightly below the threshold, and that all three critique tasks surpassed the 65% threshold only at LP Level 4.

Empirical fit of summary items to the summary learning progression

van Rijn et al. (2014) classified student argument LP levels using only argumentation items but also found that the summary items loaded on the same dimension. Interestingly, performance was worse on summary than argument items, with students typically performing about one level lower on Summary LP items. (a) Items that measured opinion versus fact and main idea versus detail were targeted at Summary LP Level 1, but the 65% threshold (seven of nine items correct) was achieved only by students at Argument LP Level 2 (see Fig. 4). (b) Items measuring the ability to identify major supporting points targeted Summary LP Level 2. As Fig. 5 demonstrates, the 65% threshold (two of three items correct) was reached by students at Argument LP Level 3. (c) Summary writing items were targeted at Summary LP Level 3. As shown in Fig. 6, Level 3 items fell well below the 65% threshold for students at Argument LP Level 3 and consistently exceeded it only at LP Level 4. (d) Finally, Summary LP Level 4 targeted items measuring the ability to recognize plagiarized text and inaccurate statements in summaries. But Summary Level 4 items did not achieve the 65% threshold for two of five items among students at argument LP Level 4 (Fig. 7). The Ban Ads plagiarism item remained difficult even at that level. (Since correct answers for summary review items were not assigned to specific categories during form design, Social Networking did not have an accuracy item.)

Research Question 2

Based upon our argument framework, each lead-in task should contribute unique variance to predicting essay performance. To evaluate this claim, we ran multiple linear regressions to predict total essay scores, entering scores on all lead-in tasks simultaneously. We considered an alternate model that added grade level after other predictors, but grade level was not significant for any of the three forms. Table 3 shows the results. Associations between the lead-in tasks and the argument writing task was moderate, with correlations between .56 and .58, and adjusted R^2 between .31 and .34. All four predictors were significant at $p < .01$. Overall, the strongest predictor was performance on the summary writing task, with standardized weights ranging between .40 and .48. The next strongest predictor was performance on the critique writing task, with standardized weights between .30 and .46. The standardized weight of this predictor was somewhat higher for Social Networking than for the other two forms. The weakest predictors of essay score were the two reading tasks. Standardized weights for the argument analysis task ranged between .15 and .20, and standardized weights for the summary reading task ranged between .12 and .23. These results indicated that all four lead-in tasks contributed unique variance to the prediction of argument writing scores.

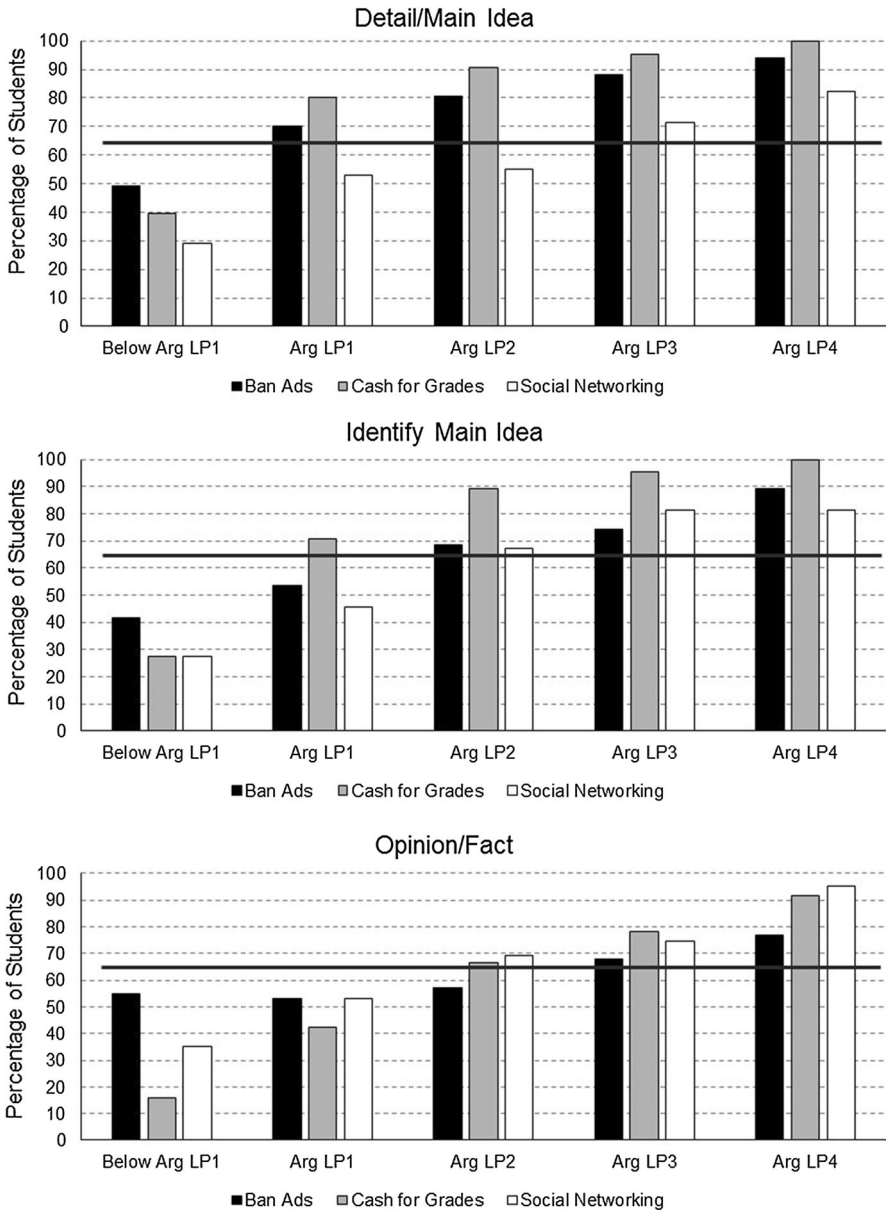


Fig. 4 Percentage correct for summary LP Level 1 Items by argument LP level. *Arg LP* Argument learning progression. The black line indicates the 65% threshold

Research Question 3

We performed latent regression IRT analysis to determine how student ability level was associated with eleven selected predictors. Eight of these were product (i.e.,

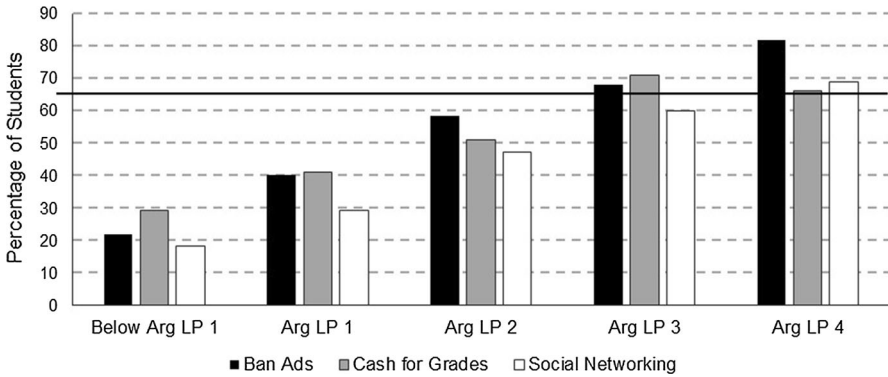


Fig. 5 Percentage of students who received full credit for summarization LP Level 2 Items by argumentation LP level. 65% threshold is indicated by black line. *Arg LP* Argument Learning progression. The black line indicates the 65% threshold

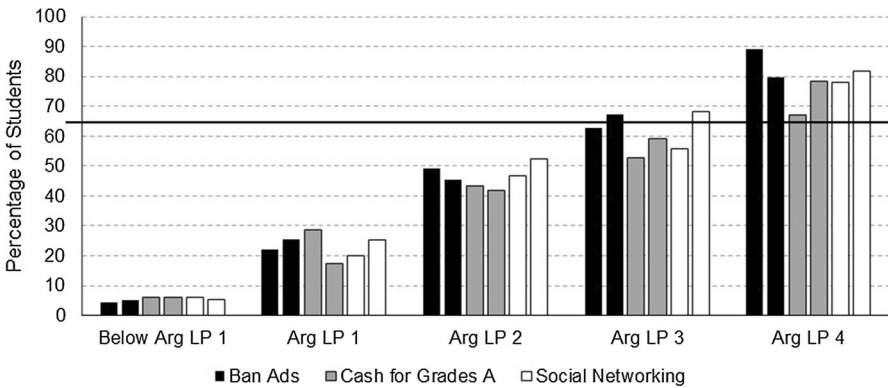


Fig. 6 Percentage of students who wrote fully or partially successful summaries (summarization LP Level 3) by argumentation LP level. *Arg LP* Argument learning progression. The black line indicates the 65% threshold

linguistic) features derived from e-rater[®], and three were process (i.e., fluency) features derived from the keystroke log. For each form, we specified a unidimensional IRT model in which ability was characterized by the seven polytomously scored tasks, as in van Rijn et al. (2014). We then compared the performance of this model with and without these predictors. Table 4 shows the results. We used the Bayesian information criterion (BIC) and IRT reliability coefficients to evaluate the contrasting models and, since the models were nested, we performed a likelihood ratio test. The models with product and process features as predictors had better relative fit (i.e., lower BIC values and significant likelihood ratio tests) and substantially higher IRT reliabilities than the models without product and process predictors. These were relatively large effects since (for example), making use of the Spearman-Brown formula, the Ban Ads form would need to be lengthened by 67% for its reliability to increase from .80 to .87.

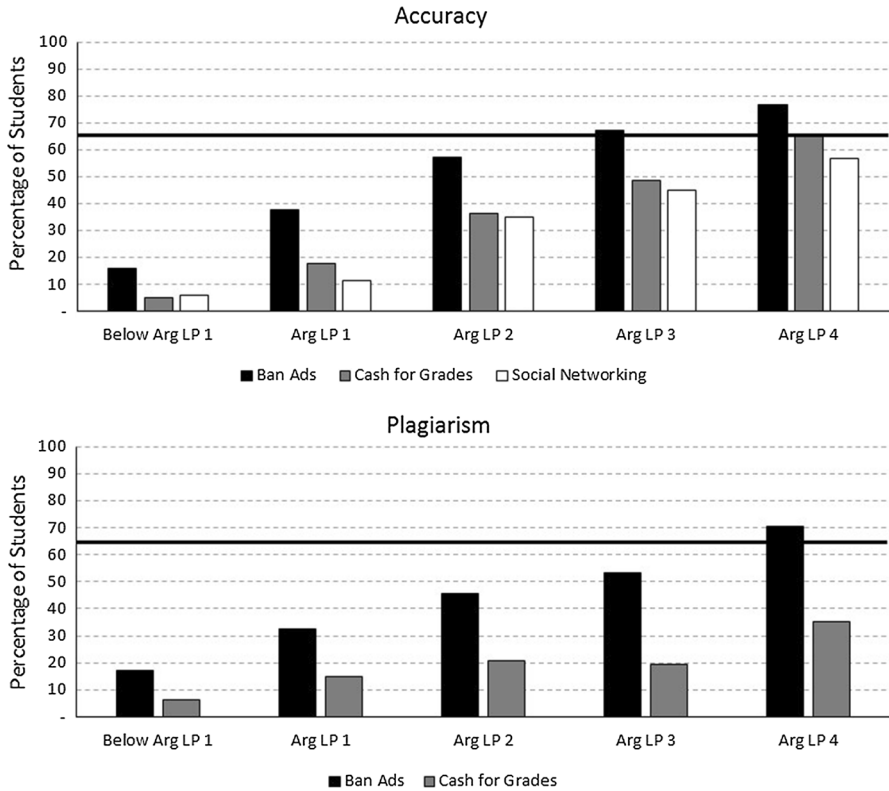


Fig. 7 Percentage of students who correctly answered summarization LP Level 4 items focusing on accuracy and recognition of plagiarism, by argumentation LP level. *Arg LP* Argument learning progression. The black line indicates the 65% threshold

Table 3 Multiple linear regressions predicting argument writing performance from supporting skills

	Ban Ads	Cash for grades	Social networking
Summary reading	.23**	.19**	.12**
Summary writing	.42**	.48**	.40**
Argument analysis	.15**	.16**	.20**
Critique writing	.30**	.30**	.46**
<i>r</i>	.56	.56	.58
Adjusted R ²	.31	.31	.34
Root MSE	1.65	1.58	1.57
F	82.14**	114.3**	119.3**

Data cells are in the format standardized coefficient (root mean square error). Ban Ads n = 711, Cash for grades n = 1016, Social networking n = 933

** = $p < .01$

Table 4 Latent regression models predicting English language arts ability level on the IRT Scale as measured by the three scenario-based assessments

	<i>n</i>	Model without predictors		Model with Predictors		Likelihood ratio test <i>G</i> ²
		BIC	IRT reliability	BIC	IRT reliability	
Ban Ads	711	13,723	.80	13,233	.87	562.1***
Cash for grades	1016	18,451	.80	17,708	.88	818.5***
Social networking	933	17,554	.82	16,843	.88	786.0***

df = 11

IRT Item response theory, *BIC* Bayesian information criterion, *G*² likelihood ratio Chi-squared

As shown in Table 5, fluency features (process features including the number of bursts, mean log length of bursts, and burst pacing) contributed statistical significance to the model, with positive weights indicating that more fluent writers were generally more able. The number of bursts had standardized weights between .18 and .32, and mean log-burst length had standardized weights between .21 and .39. Burst-pacing standardized weights ranged between .11 and .32. Avoidance of grammar, usage, mechanical, and stylistic errors were also consistently significant predictors ($p < .01$), except for usage errors in Ban Ads. Mechanics had standardized weights between .40 and .51; grammar, between .27 and .36; and style, between .25 and .37. Idiomaticity was never statistically significant.

Finally, features measuring language sophistication were statistically significant predictors. Word length had significant weights in Cash for Grades (.31) and Social Networking (.38). Word frequency had significant weights of .33 in Ban Ads and .23 in Social Networking. Syntactic variety was statistically significant across all models, with weights between .27 and .51. Overall, students who produced higher-quality essays were more fluent, produced fewer errors, and had more varied linguistic expression than did students who wrote lower-quality essays.

Study 1 discussion

Research Question 1

As the above analysis indicates, the theoretical design mapped well onto performance patterns at the individual item level (not just tasks, as in previous studies, where performance on SR items was pooled), or in the case of polytomous items, individual options or option choices. Students in our study appeared to be, in general, one level more advanced on the argument than on the summary learning progression, which might be a function of the nature of the summary required. The argument and summary LP levels can therefore be interpreted as theoretically-motivated performance patterns that reflect written argument skill levels, with typical performance profiles varying by level on a *combined* progression, as follows. This progression begins below Level 1 (since some students failed to accurately

Table 5 Product and process feature coefficients from latent regression IRT analysis

	Number of bursts	Mean log burst length	Burst pacing	Avoidance of grammar errors	Avoidance of usage errors	Avoidance of mechanics errors	Avoidance of stylistic errors	Idiomacity of language	Syntactic variety	Word length	Lower median word frequency
Ban Ads	.32**	.25**	.22**	.28**	.13	.40**	.25**	-.05	.51**	.01	.33**
Cash for grades	.18*	.21**	.11*	.27**	.23**	.51**	.37**	.09	.50**	.31**	.10
Social networking	.44**	.39**	.32**	.36**	.23**	.40**	.26**	-.02	.27**	.38**	.23**

* $p < .05$; ** $p < .01$; *** $p < .00$

answer even Level 1 items) and does not include Level 5 (because Level 5 items are more appropriate for high school- or college-level populations).

Below Level 1

Students could classify some statements as being for or against a position and could often recognize when evidence strengthened an argument, without performing consistently on either task. They had difficulty recognizing main ideas and thus in summarizing source articles that contained arguments; in fact, they were typically unsuccessful on all the summary tasks. They also had difficulty writing full argument essays, recognizing evidence that weakened an argument, and creating a critique.

Level 1

Students could consistently classify statements as supporting or opposing a position and usually recognize when evidence strengthened an argument. Their argument essays did little more than state their opinion. Their responses to the written critique task were also minimal, though they sometimes identified flaws in arguments they were asked to critique. They often had difficulty performing summary tasks, even on relatively easy items.

Level 2

Students could consistently recognize when evidence weakened (not just strengthened) an argument and often produced argument essays with a thesis and multiple supporting reasons, though their arguments were not always elaborated well. They had difficulty writing argument critiques, yet they often recognized multiple weaknesses in a text. They could perform simple summary tasks on written argument, such as distinguishing opinion from statements of fact, distinguishing main ideas from details, and identifying the main idea of an article.

Level 3

Students could identify main and supporting points in a written argument and express them in summary form, but their summaries were sometimes incomplete, inaccurate information, or partially plagiarized. They could analyze written arguments and write fully developed arguments of their own, but often encountered difficulty writing a formal critique or conducting a fine-grained analysis of textual wording and content.

Level 4

Students could analyze, summarize, and critique written arguments effectively, and produce well-developed arguments. Additionally, they were often able to analyze a text for inaccurate or plagiarized statements.

These patterns were generally consistent across forms, even at the item level, and suggest hypotheses about appropriate instructional interventions. For example, below Level 1, instruction might include oral debate to build intuitions about when evidence strengthened, undermined, or was irrelevant to an argument. At Level 1, the instructional focus might include an explicit mapping of argument structure and then using that information to construct a summary or elaborate an essay by constructing multiple supporting arguments. To reach Level 2, students would also need to become much more effective at recognizing and evaluating counterarguments. To reach Level 3, they would need to learn how to represent the full hierarchical structure of a text; analyze an argument in terms of claims, reasons, and evidence; and produce essays that develop these elements fully. Reaching Level 4 requires opportunities to practice fine-grained analysis and evaluation of their own and others' textual arguments. These hypotheses are consistent with what has been reported in the literature.

It is worth noting that performance on one of the three essay tasks did not exceed 65% until students reached Argument LP Level 4. This may reflect the relative difficulty of argument essay writing in the grade levels examined (grades 7–9) as well as a generally low level of writing proficiency in 8th grade (NCES, 2012).

Unexpectedly, summary tasks were more difficult than argument tasks. However, students were required to summarize argument texts, not texts written in easier genres with which they may have been more familiar. These results suggest that unless students can analyze basic argument structures (e.g., thesis, claims, evidence, and rebuttal), they may not understand the structure of arguments in written texts well enough to summarize the content accurately. An alternative explanation is that students are not consistently taught how to write quality text summaries. Summary writing requires a disciplined approach to evaluating and organizing essential ideas in a text—a task closely aligned with formation of a coherent mental model of text content (Wang, Sabatini, O'Reilly, & Feng, 2017).

Research Question 2

Our results indicate that each of the major reading and writing tasks built into our scenario-based assessment design contributed significantly to essay score prediction, though there was also significant unexplained variance in writing scores that may have been due to the fact that only the essay task required the writer to draft a multiple-paragraph text. Since performance was aligned with LP levels (as demonstrated under Research Question 1) and performance on the lead-in tasks was moderately associated with levels of writing performance (as demonstrated under Research Question 2), these results suggest that the pattern of performance on the entire scenario-based assessment provided a reasonable profile of student competency that might help identify where students are likely to do well (or run into trouble) during an extended process of understanding, analyzing, and producing textual arguments. That is, our scenario-based assessment design appeared to measure a complex set of skills that contribute to student success in handling textual argument and provided richer information than could be obtained from a single performance measure, such as the final essay task alone. Our results also indicate

that written argument instruction may need to address a wide range of summary and argument skills, since deficits in a contributing skill can derail the overall performance.

The relative strength of summary writing as a predictor for essay writing performance was, however, striking and somewhat unexpected. This pattern could be accounted for as follows. First, as argued above, successful performance on the summary tasks appeared to presuppose progress in understanding and analyzing the structures of argument. Second, summary writing was a productive task and thus might also be constrained by general writing fluency. Third, weakness in summary writing is an indicator of weak comprehension, i.e., a weak mental model formation of text content, which impacts subsequent argument writing performance (Gil, Braten, Vidal-Abarca, & Stromso, 2010).

Research Question 3

Higher ability students in written argument might be expected, on average, to demonstrate stronger basic literacy skills, such as control over spelling and grammar, fluent keyboarding skills, and mastery of oral and academic language. Conversely, when students attempt a written argument task, underlying deficits in fundamental literacy skills might reduce the working memory available to generate, evaluate, or analyze arguments. Our results were consistent with this theoretical account, since linguistic- and process-based features of student essays were predictors of overall ability in written argument as measured by the full assessment.

Further implications

Our results also indicated that useful information about fundamental literacy skills could be extracted from a single writing sample, if appropriate automated writing evaluation tools are used. Many students may not perform well on written argument tasks due to difficulties in verbal ability, keyboarding fluency, or control of standard written English—all of which could be efficiently identified with automated methods using digitally administered writing tasks. In effect, the process (fluency) and product (linguistic) features provided a secondary skill profile that supplements information provided by the lead-in tasks.

Study 2

Using data from reported by Zhang, Van Rijn, Deane, and Bennett (2017), Study 2 investigated whether changes in the SBA structure yielded better information about the performance of lower-performing students. We wanted to know how the scenario sequence affected student writing processes or the content and quality of student essays. We hypothesized that the chief effect of the scenario structure was to encourage students to read source articles and analyze the issue *before* attempting the essay task so that they would be better prepared, acquire deeper knowledge, and/or have activated more relevant prior knowledge during the writing process (Gil

et al., 2010). Such an effect would be beneficial if it reduced the cognitive load of writing, giving students a better opportunity to demonstrate how well they could write when they were relatively well-prepared.

Zhang et al. (2017) prepared four variants on our written argument assessment design, manipulating the order of essay and lead-in tasks and changing the content of the lead-in tasks in relation to the topic of the essay task. Students were randomly assigned to one of the four forms. Of particular relevance to the current paper are comparisons of the *intact* condition (the form in which the lead-in tasks preceded the essay task and focused on the same topic) to the two forms in the *reversed-order* condition, where the essay was written first, without the support provided by the lead-in tasks. They found that, compared to students in the reversed order condition, those receiving the intact SBA achieved similar essay and total test scores but wrote shorter essays, spent less time composing, and produced essay scores that had a stronger correlation to time spent composing, which suggests a more direct connection between effort and quality of essay.

One result from Zhang et al. (2017) was particularly striking: their observation that students wrote longer essays in the reversed-order condition, though without significant differences in essay scores. It has frequently been observed that longer essays tend to obtain higher scores (Breland, Camp, Jones, Morris, & Rock, 1987), yet in this case, differences in average essay length were not correlated with scores. We wanted to understand how student writing behavior differed between the two conditions. One explanation for the increased length might be that students in the reversed-order condition used more direct quotations or paraphrases of the sources. An alternative explanation could be that students' original language was wordier, reflecting less efficient, less focused text-generation processes. Either mechanism could increase text length without yielding a parallel increase in score. However, it would be far more interesting if the intact order induced greater efficiency in students' generation of original language, since that behavior would suggest that the lead-in task was having a deep effect on student writing processes. This question could be tested in part by examining the relation between original and source-based text in student essays under the two conditions. If students in the reversed-order condition used a greater proportion of original text compared to students in the intact, scenario-based order, the intact order might be inducing more efficient text generation.

The issue of writing fluency and efficiency could also be addressed by examining students' keyboarding patterns under the two conditions. While keyboarding patterns provide relatively little information about some aspects of the writing process, such as planning, they provide direct evidence about how a student creates a piece of writing online and, for that reason, may have diagnostic and instructional value. In a follow-up study using the same dataset, Zhang et al. (2017) examined differences in students' word processing patterns in the intact and reversed-order conditions, using a fine-grained set of features that classified keyboarding pauses by their length and linguistic properties. The largest effect they observed involved features associated with typing fluency (e.g., the speed and consistency of individual keystrokes). These features accounted for more of the variance in essay score in the reversed-order condition. Other features also displayed statistically significant differences between the intact and reversed-order conditions, most notably the

length of long pauses between bursts of keyboarding actions. Long pauses were generally even longer in the reversed condition, where they were more strongly associated with the quality of essay content (as measured by one of the two rubrics used to score the essays).

Zhang et al. (2017) focused on factor analysis of keystroke features collected as part of the study reported Zhang et al. (2017). However, the patterns they observed can be interpreted cognitively in terms of the concept of *burst*, i.e., rapid keystroke sequences delimited by long pauses (Chenoweth & Hayes, 2001). As various authors have argued (Alves & Limpo, 2015; Fayol, 1999; Schilperoord, 2002), burst length is highly sensitive to working-memory demands during writing. Component writing processes (idea generation, translation of ideas into language, transcription, and executive control or monitoring processes) compete for limited working-memory resources during writing (Kellogg, 1996; McCutchen, 1996). Any source of increased working-memory load during writing is likely to increase the frequency and duration of pauses, while decreasing burst length. Since the intact SBA order was intended to reduce the burden imposed by the need to read, analyze, and remember material from the source articles during the writing process, we would expect greater writing fluency in the intact order. Conversely, in the reversed order, the need to coordinate reading and analysis of source articles with essay writing should increase working-memory load. We thus expect to observe longer pauses (and shorter bursts) when students were writing without the support provided in the intact order SBA.

In Study 2, we examined the following research question:

1. Does the scenario-based structure of the assessment change the way students write in ways that might reflect better support for low-performing students? In particular, does completing the essay after a series of lead-in tasks on the same topic
 - (a) result in students' producing essays with proportionately less original text compared to their essays written without any previous exposure to the content, and
 - (b) result in longer bursts of text production?

Question (1a) examines whether the support provided by the scenario resulted in more efficient text production (with students achieving comparable scores from shorter essays). Question (1b) examines whether the scenario structure reduced cognitive load during the writing task, resulting in greater writing fluency.

Method

For the current study, data were analyzed from three of the four forms¹ studied in 2014 by Zhang et al. (2017): (a) the original Ban Ads with its order intact, (b) the Ban Ads form with the order of sessions reversed, and (c) a form in which students

¹ Since the fourth form did not meet the requirements of the current analysis, it is not described.

completed first the Ban Ads essay task and then the Cash for Grades lead-in tasks (with unrelated content). Students were assigned randomly to the variant forms within classrooms. Each form was administered in two 45-min sessions taken on different days with little intervening time. Internal consistency reliabilities were similar across forms, with coefficient alphas between .80 and .84. Correlations of the total test score with teachers' ratings of students' writing skill were also comparable across forms, from $r = .50$ to $r = .54$.

The sample was comprised of 1089 8th-grade students from eight schools across eight U.S. states, with slightly more males (573) than females (516). With regard to race/ethnicity, 884 were Caucasian, 104 African-American, 48 Hispanic, 30 Asian, and 23 other. In terms of English language proficiency, 903 were classified as proficient in English, 175 as English learners, and 11 were English learners reclassified as fluent.

Data analysis

For Research Question 1a, we examined essay content using a combination of manual and semi-automated methods to identify any significant test-form-based differences between students' use of original vocabulary versus their use of vocabulary drawn from the source articles (which were available for students to read at all times under both conditions). We hypothesized that essays written to the intact order form would employ topic-specific, infrequent words from the source articles with greater relative frequency, but that the increased length of essays written in the two reversed conditions would primarily reflect an increase in the proportion of original language, as evidenced by the production of infrequent words that were not in the sources.

To test this hypothesis, we created two corpora. One corpus comprised words from essays written in response to the intact order form. The other consisted of words from essays written in response to either of the two reversed forms. To minimize loss of data due to misspelling, we standardized spelling using Microsoft Word's spell-check feature and created a database of corrected words. Each word in the database was associated with the following attributes: the form in which it appeared, whether it was in the source articles, and its Standardized Frequency Index (SFI) in the Touchstone Applied Scientific Associates (TASA) corpus (Zeno, Ivens, Millard, & Duvvuri, 1995).

To measure students' source use, we examined the extent to which infrequent words from the source articles appeared in student essays. We defined infrequent words as those with an SFI less than 55, using SFIs derived from the TASA corpus (Zeno et al., 1995). This SFI criterion corresponded to words like *celebrities*, *forum*, *playmates*, *sponsor*, *research*, *batteries*, *Swedish*, *Netherlands*, *Belgium*, *evaluate*, *media*, *childhood*, *psychological*, *commission*, *candy*, and *habits*. We excluded a few words with TASA SFI values less than 55 because they were morphological variants or synonyms of prompt words, such as *commercials*, *advertisements*, and *advertisers*.

We then analyzed the relative frequency of words drawn from the source articles across the two conditions, intact and reversed. Chi-squared tests were run to contrast

the distribution of infrequent source-based words versus equally infrequent original words across the two conditions. Standardized residuals were examined for each cell in the comparison, which indicated the extent to which the relative frequency of words in that cell differed from what would be expected by chance. Essay word counts were also computed and compared across cells.

For Research Question 1b, we presumed that students who had already read and analyzed the arguments before they started writing would know more about the topic, which should have reduced their working-memory load and lessened the need to return to the source articles to find useful information. We therefore hypothesized that students would have produced longer, more variable-length bursts in the original intact order, but shorter and less variable (but more frequent) bursts of text production in the reversed condition. We therefore examined the frequency, average length, and standard deviation of text-production bursts during each student's writing process. As in Study 1, this information was captured online using in the burst definition from Almond et al. (2012).²

Results

Research Question 1a

A Chi-squared test revealed a significant relation between form administered and use of infrequent words that did or did not appear in the source articles ($X^2(1) = 70.54, p < .001$). In the intact condition, 49.8% of the words in the student essays did not appear in the source articles. By contrast, in the reversed condition, 57.3% of the infrequent words in student essays did not appear in the source articles (see Table 5). As previously observed by Zhang et al. 2017, a one-way ANOVA indicated a statistically significant difference in the total number of words per essay between the original and reversed conditions $F(1741) = 18.39, p < .001$. This contrast corresponded to students producing roughly the same number of infrequent words from the source articles per essay in both conditions (8.9 vs. 8.6 words per essay), but significantly fewer original infrequent words in the intact condition than in the reversed-order condition (8.9 vs. 11.7 words per essay). A comparison of the standardized residuals (Agresti, 2007) confirms this result (again, see Table 6), with the standardized residual for each cell being well above 2, the value required to show significant variance from the expected proportions.

Research Question 1b

As part of the current study, we conducted an analysis of variance to examine the effect of form on the distribution of typing bursts during the writing process. The results were statistically significant ($F(2987) = 83.23, p < .001$). Additionally, the mean log length of those bursts (in characters) was significant ($F(2921) = 6.39,$

² Due to issues that arose during the web-based data collection, the writing log data was not successfully collected for about 7% of the essays. The data loss does not appear to be systematic and hence should have minimal impact on the results.

Table 6 Total number of infrequent words found across essays

Test condition	Infrequent words absent from the source articles	Infrequent words present in the source articles	Total
Original	2123 (− 4.7)	2136 (5.2)	4259
Reversed	5874 (3.0)	4348 (− 3.4)	10,195

Words were judged as infrequent when their TASA SFI ≤ 55 . TASA SFI = A measurement of word frequency developed Touchstone Applied Science Associates. For more information, see Zeno et al. (1995)

$p < .001$), as was the standard deviation of log length in characters ($F(2915) = 7.37$, $p < .002$). Students produced fewer bursts of typing in the original intact order, but longer average bursts, with greater variation in length (see Table 7) as compared to the reversed order conditions.

Study 2 discussion

Research Question 1a

Our results confirmed that students writing in the reversed-order condition were significantly likely to produce more original vocabulary (infrequent words that were *not* drawn from the sources) than students writing in the intact order. By comparison, infrequent words drawn from the sources occurred at roughly similar rates.

The generally observed positive relationship between document length and the ratio of types (total number of different words) to tokens (total number of words) would ordinarily imply that longer essays will contain a larger number of less-frequent words (Baaijen, 2001). The fact that this effect appears only in students' original words supports the conclusion that the difference between the reverse-order and intact-order conditions primarily affected students' original writing, not their use of material extracted from the sources.

And yet, as Zhang et al. previously reported, there was no significant difference in essay score between the intact and reversed-order conditions, even though essays written in the reversed-order condition were significantly longer. It thus seems likely that students in the reversed-order produced roughly comparable content to that by students writing in the intact condition, but expressed that content less concisely. We hypothesize that the intact-order condition encouraged deeper processing of information from the source texts, which in turn enabled students to express their ideas more efficiently.

Research Question 1b

We expected, based on the psycholinguistic literature, that the original scenario order would reduce working-memory load and increase writing fluency, enabling writers to shift more attention from basic writing processes to meet other task

Table 7 Comparison of burst features for original and reversed order

Form	Number of essays	Mean number of bursts	Mean log length of bursts	Standard deviation of log length of bursts
Intact order	218	113.82 (72.61)	1.62 (.37)	.99 (.14)
Reversed order	474	155.97 (86.86)	1.52 (.35)	.95 (.14)

Standard deviations are shown in parentheses. Due to a system error, a small number of process logs were not successfully recorded resulting in some reduction in sample sizes

demands. In particular, we hypothesized that the intact sequence, by requiring students to read and analyze source content before they write their essay, would increase and activate students' prior topic knowledge, reducing the cognitive load of reading and remembering information about the topic during writing. Conversely, we expected students writing in the reversed order condition would have less working memory and therefore write less fluently, since they would have to devote additional time and attention to the preparatory tasks of reading the source articles, analyzing the arguments, and planning their essays. Our results were consistent with these expectations. Students writing in the reversed conditions produced shorter (and less variable) bursts of text production. The simplest explanation for this pattern is that students who wrote essays under the reversed-order condition were operating under a heavier cognitive load.

However, the fact that students produced longer essays (and therefore more bursts) in the reversed order is less easily explained. One possibility is that students in both conditions were working toward a specific target—a minimum level of content elaboration that they considered sufficient for the task—and stopped writing as soon as that target level was reached. This interpretation suggests that students who wrote more fluently and efficiently produced fewer wasted words, ending up writing shorter essays that reached the same quality standards sooner.

Conclusion

In recent years, educators, researchers and policy makers have called for an updated, more demanding reading/writing construct (CCSSO & NGA, 2010; Goldman et al., 2016). However, scores on traditional summative assessments often have limited instructional value, revealing only how poorly many students perform. Accordingly, this investigation sought to introduce and evaluate a scenario-based assessment (SBA) designed to measure the argument construct while also providing more instructionally useful information.

Overall, these studies provide additional validity evidence to support the use of an SBA of written argument skills. In particular, the results of Study 1 indicate that the argument and summary LPs that formed the basis for the SBA design helped provide a rich characterization of student performance patterns at the LP levels.

Each lead-in task contributed unique variance toward predicting essay score, indicating its potential usefulness as a component index. Study 1 also demonstrated that independent information on student performance can be obtained from automatically computed linguistic measures of student writing processes. Combining this information with LP level assignments provided richer and more reliable descriptions of student performance. The results of Study 2 suggest that the scenario sequence and topically relevant lead-in tasks had the intended effect of supporting students so that they could better demonstrate their argument writing skills in the essay task.

The information provided by this kind of SBA design may support instruction by helping teachers build on student strengths and address their weaknesses. Because SBAs provide rich information about what lower-performing students know and can do, our design may help teachers determine whether weaknesses in written argument skills are due to issues in fundamental literacy, in specific argument skills, or both. Students who perform well on the lead-in tasks but show weaknesses on features captured by automated essay scoring have fundamentally different instructional needs from students who produce clean, grammatically correct text reasonably fluently but demonstrate a low level of performance on the argument lead-in tasks.

In particular, our results indicate that students below argument LP Level 3 were not effective at tasks that required explicit, metacognitive representation of argument structure. This suggests that there may be considerable value in explicit instruction designed to build up students' ability to reason explicitly about argument. Even students with strong fundamental literacy skills are likely to struggle when asked to read and summarize argument texts (or to write extended arguments of their own) before they have achieved explicit metacognitive awareness of argument.

Most of the evidence that placed students below Level 3 in the Argument and Summary LPs came from selected-response and summary items, which can be administered in much less time than is required by the entire SBA. In an instructional context, it may therefore be useful for teachers to identify at-risk students using a small diagnostic battery focused on LP Levels 1 and 2 and then devote significant effort to helping these students learn to analyze and generate ideas for argument before subjecting them to the frustration and likely demotivating effect of attempting writing tasks they are not yet prepared to address.

Several limitations of our two studies should be noted. Study 1 had very few students from minority groups and drew the majority of its sample from a single western state. Study 2 had a larger representation of low-SES students (about 25%), but neither study had a large proportion of English learners. Also, the current study and the data collections on which it was based were almost exclusively cross-sectional; we do not yet know how specific instruction or more general developmental processes might affect student performance on this kind of SBA design.

Our results suggest several directions for future research. One of considerable interest is characterizing more precisely how student essays differed qualitatively between the intact and reversed conditions in Study 2. We determined that the

differences primarily occurred in the students' production of original text (words not in the source articles). Since there was no detectable effect on score distributions, it seems reasonable to interpret this effect as one in which essays in the reversed condition were wordier, without a corresponding increase in relevant content. We were not, however, able to characterize exactly what students in the reversed condition did that made their text less succinct than essays written in the original (intact) SBA order. In future studies, it will be critical to examine developmental patterns in greater detail and link the patterns we observe in SBAs more closely with student learning. The structure of argument SBAs may yield a more detailed picture of longitudinal changes in student argument skills and help identify circumstances in which students become better able to analyze other people's arguments and produce stronger arguments on their own. It may also be useful to explore in greater depth the extent to which improvements in performance on the argument and summary tasks are driven by students' awareness of argument structure rather than variables such as reading and writing fluency, which affect most literacy tasks.

References

- Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: Wiley.
- Almond, R., Deane, P., Quinlan, T., & Wagner, M. (2012). A preliminary analysis of keystroke log data from a timed writing task. *ETS Research Report Series, 2012(2)*. Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2012.tb02305.x>.
- Alves, R. A., & Limpo, T. (2015). Progress in written language bursts, pauses, transcription, and written composition across schooling. *Scientific Studies of Reading, 19(5)*, 374–391. <https://doi.org/10.1080/10888438.2015.1059838>.
- Attali, Y., & Burstein, J. (2005). Automated essay scoring with E-Rater v. 2.0. *ETS Research Report Series, 2004(2)*. Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2004.tb01972.x>.
- Baaijen, H. R. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.
- Bennett, R. E. (2011). CBAL: Results from piloting innovative K-12 assessments. *ETS Research Report Series 2011(1)*. Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2011.tb02259.x>.
- Bennett, R. E., Deane, P., & van Rijn, P. W. (2016). From cognitive-domain theory to assessment practice. *Educational Psychologist, 51(1)*, 1–26. <https://doi.org/10.1080/00461520.2016.1141683>.
- Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. R. (1987). *Assessing writing skill* (College Board Research Report No. 11). New York: College Entrance Examination Board.
- Brem, S. (2000). Explanation and evidence in informal argument. *Cognitive Science, 24(4)*, 573–604. https://doi.org/10.1207/s15516709cog2404_2.
- CCSSO, & NGA. (2010). *Common core state standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC.
- Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication, 18(1)*, 99–118. <https://doi.org/10.1177/0741088301018001004>.
- Deane, P., Fowles, M., Baldwin, D., & Persky, H. (2011). The CBAL summative writing assessment: A draft eighth-grade design. *ETS Research Memorandum Series* (Report No. RM-11-01). Princeton, NJ: Educational Testing Service. <https://sharepoint.etslan.org/rd/rreports/RR/RM-11-01.pdf>.
- Deane, P., Sabatini, J., Feng, G., Sparks, J., Song, Y., Fowles, M., et al. (2015). Key practices in the English language arts (ELA): Linking learning theory, assessment, and instruction. *ETS Research Report Series*. <https://doi.org/10.1002/ets2.12063>.

- Deane, P., & Song, Y. (2014). A case study in principled assessment design: Designing assessments to measure and support the development of argumentative reading and writing skills. *Educativa Psicología*, 20(2), 99–108. <https://doi.org/10.1016/j.pse.2014.10.001>.
- Deane, P., & Song, Y. (2015). The key practice, discuss and debate ideas: Conceptual framework, literature review, and provisional learning progressions for argumentation. *ETS Research Report Series*, 2015(2). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/ets2.12079>.
- Fayol, M. (1999). From on-line management problems to strategies in written composition. In M. Torrance & G. Jeffery (Eds.), *The cognitive demands of writing: Processing capacity and working memory effects in text production* (pp. 15–23). Amsterdam: Amsterdam University Press.
- Ferrari, M., Bouffard, T., & Rainville, L. (1998). What makes a good writer? Differences in good and poor writers' self-regulation of writing. *Instructional Science*, 26(6), 473–488. <https://doi.org/10.1023/A:1003202412203>.
- Ferretti, R. P., Lewis, W. E., & Andrews-Weckerly, S. (2009). Do goals affect the structure of students' argumentative writing strategies? *Journal of Educational Psychology*, 101(3), 577–589. <https://doi.org/10.1037/a0014702>.
- Fu, J., Chung, S., & Wise, M. (2013). Dimensionality analyses of CBAL Writing tests. *ETS Research Report Series*, 2013(1). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2013.tb02317.x>.
- Fu, J., & Wise, M. (2012). Statistical report of 2011 CBAL multistate administration of reading and writing tests. *ETS Research Report Series*, 2012(2). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2012.tb02306.x>.
- Gil, L., Braten, I., Vidal-Abarca, E., & Stromso, H. I. (2010). Summary versus argument tasks when working with multiple documents: Which is better for whom? *Contemporary Educational Psychology*, 35(3), 157–173. <https://doi.org/10.1016/j.cedpsych.2009.11.002>.
- Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.
- Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M. A., Greenleaf, C., et al. (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist*, 51(2), 219–246. <https://doi.org/10.1080/00461520.2016.1168741>.
- Hayes, J. R. (2012). Evidence from language bursts, revision, and transcription for translation and its relation to other writing processes. In M. Fayol, D. Alamargot, & V. W. Berninger (Eds.), *Translation of thought to written text while composing*. New York: Psychology Press.
- Hillocks, G. (2002). *The testing trap: How state writing assessments control learning*. New York: Teachers College Press.
- Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57–71). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Klaczynski, P. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development*, 71(5), 1347–1366. <http://www.jstor.org/stable/1131978>.
- Kuhn, D. (1991). *The skills of argument*. Cambridge: Cambridge University Press.
- Kuhn, D., & Crowell, A. (2011). Dialogic argumentation as a vehicle for developing young adolescents' thinking. *Psychological Science*, 22(4), 545–552. <https://doi.org/10.1177/0956797611402512>.
- Kuhn, D., Shaw, V., & Felton, M. (1997). Effects of dyadic interaction on argumentative reasoning. *Cognition and Instruction*, 15(3), 287–315. <http://www.jstor.org/stable/3233770>.
- Kuhn, D., & Udell, W. (2007). Coordinating own and other perspectives in argument. *Thinking & Reasoning*, 13(2), 90–104. <https://doi.org/10.1080/13546780600625447>.
- Leitão, S. (2003). Evaluating and selecting counterarguments: Studies of children's rhetorical awareness. *Written Communication*, 20(3), 269–306. <https://doi.org/10.1177/0741088303257507>.
- Levy, C. M. (2013). *The science of writing: Theories, methods, individual differences and applications*. New York: Routledge.
- Mayweg-Paus, E., Macagno, F., & Kuhn, D. (2016). Developing argumentation strategies in electronic dialogs: Is modeling effective? *Discourse Processes*, 53(4), 280–297. <https://doi.org/10.1080/0163853X.2015.1040323>.
- McCann, T. M. (1989). Student argumentative writing: Knowledge and ability at three grade levels. *Research in the Teaching of English*, 23(1), 62–76. <http://www.jstor.org/stable/40171288>.

- McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, 8(3), 299–325. <https://doi.org/10.1007/BF01464076>.
- NCES. (2012). The Nation's report card: Writing 2011. National Center for Education Statistics. <https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf>.
- Nussbaum, M. E., & Kardash, C. (2005). The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology*, 97(2), 157–169. <https://doi.org/10.1037/0022-0663.97.2.157>.
- O'Reilly, T., Deane, P., & Sabatini, J. (2015). Building and sharing knowledge key practice: What do you know, what don't you know, what did you Learn? *ETS Research Report Series*, 2015(2). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/ets2.12074>.
- O'Reilly, T., Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing reading comprehension assessments for reading interventions: How a theoretically motivated assessment can serve as an outcome measure. *Educational Psychology Review*, 26(3), 403–424. <https://doi.org/10.1007/s10648-014-9269-z>.
- Sabatini, J. P., Halderman, L. K., O'Reilly, T., & Weeks, J. P. (2016). Assessing comprehension in kindergarten through third grade. *Topics in Language Disorders*, 36(4), 334–355. <https://doi.org/10.1097/TLD.000000000000104>.
- Sabatini, J. P., O'Reilly, T., Halderman, L. K., & Bruce, K. (2014). Integrating scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disabilities Research & Practice*, 29(1), 36–43. <https://doi.org/10.1111/ldrp.12028>.
- Schilperoord, J. (2002). On the cognitive status of pauses during text revision. In T. Olive & C. M. Levy (Eds.), *Contemporary tools and techniques for studying writing* (pp. 61–87). Dordrecht: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-010-0468-8_4.
- Shemwell, J. T., & Furtak, E. M. (2010). Science classroom discussion as scientific argumentation: A study of conceptually rich (and poor) student talk. *Educational Assessment*, 15(3–4), 222–250. <https://doi.org/10.1080/10627197.2010.530563>.
- Van der Schoot, F. C. J. A. (2002). *The application of an IRT-based method for standard setting in a three-stage procedure*. In Paper presented at the the annual meeting of the National Council on Measurement in Education, New Orleans, April 2–4, 2002.
- van Rijn, P. W., Graf, E. A., & Deane, P. (2014). Empirical recovery of argumentation learning progressions in scenario-based assessments of English language arts. *Educative Psicologia*, 20(2), 109–115. <https://doi.org/10.1016/j.pse.2014.11.004>.
- van Rijn, P. W., & Yan-Koo, Y. (2016). Statistical results from the 2013 CBAL English Language arts multistate study: Parallel forms for argumentative writing. *ETS Research Monograph Series* (Report No. RM-16-15). Princeton, New Jersey: Educational Testing Service.
- Wang, Z., Sabatini, J. S., O'Reilly, T., & Feng, G. (2017). How individual differences interact with task demands in text processing. *Scientific Studies of Reading*, 21(2), 165–178. <https://doi.org/10.1080/10888438.2016.1276184>.
- White, E. M. (1995). An apology for the timed impromptu essay test. *College Composition and Communication*, 46(1), 30–45. <http://www.jstor.org/stable/358868?origin=JSTOR-pdf>.
- Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.
- Zhang, M., Van Rijn, P., Deane, P., & Bennett, R. E. (2017). Scenario-based assessments in writing: An experimental study. Manuscript submitted for publication.
- Zhang, M., Zou, D., Wu, A. D., Deane, P., & Li, C. (2017b). 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. Social indicators research series* (Vol. 69). Cham: Springer. https://doi.org/10.1007/978-3-319-56129-5_17.