Reading and Writing (2024) 37:103–127
https://doi.org/10.1007/s11145-022-10407-x

Assessing argumentation skills of middle school students: a learning progression approach

Yi Song[1] · Peter van Rijn[2] · Paul Deane[1] · Szu-Fu Chao[1]

Accepted: 19 December 2022 / Published online: 25 February 2023

Abstract

Argumentation skills are emphasized by the common core state standards (CCSS) and are viewed as essential for success in college, career, and life. Our project aims to develop formative assessment tasks measuring students' argumentative reading and writing skills. We used the framework of the *Discuss and Debate Ideas* key practice (describing the key argumentation skills) to guide the task development and gathered evidence of students' argumentation skills. Specifically, we designed 27 tasks targeting various argumentation skills, spanning three learning progression (LP) levels aligned with the grade level expectations for argumentation in CCSS. The goal was to evaluate the potential utility of an LP-based approach to task design in assessing argumentation skills. We conducted a study with 786 seventh- and eighth-grade students to examine how well these tasks function and identify patterns of student performance. We also examined whether student performance patterns aligned with predicted LP levels, using task progression maps derived from item response theory (IRT) models. Results suggested that the majority of tasks were reliable, and that the LP-based tasks were significantly related to students' reading skills. Student LP performance was most strongly correlated with reading comprehension as measured by the RISE assessment, and was also significantly associated with foundational reading skills, such as word recognition, decoding, and vocabulary. However, some lower-level LP tasks appeared to be unexpectedly difficult. We found evidence that such factors as topic, task format, and scoring could have contributed to item difficulty and affected student task performance. Implications for future LP research are provided.

Keywords Argumentation · Learning progression · Assessment · Task design · English language arts

✉ Yi Song
ysong@ets.org

[1] Educational Testing Service, Princeton, NJ 08550, USA

[2] ETS Global, Amsterdam, The Netherlands

Introduction

Argumentation skills are essential to academic success (Graff, 2003) and play a prominent role in standards-based reform initiatives such as the Common Core State Standards (CCSS; Council of Chief State School Officers, 2010). The CCSS require students to learn to construct and evaluate arguments in the English Language Arts (ELA), history, science, and math. For example, the College and Career Readiness anchor standards from the ELA writing expects students to "write arguments to support claims in an analysis of substantive topics or texts using valid reasoning and relevant and sufficient evidence".[1] Engaging in evidence-based argumentation with multiple sources of information also contributes to the development of disciplinary literacy that is needed for the development of expertise in subject areas (Ferretti & De La Paz, 2011; Goldman et al., 2016; Shanahan & Shanahan, 2008).

Because argumentation requires higher order thinking processes, it is a challenging skill to learn and to teach (Gilbert & Graham, 2010; Kiuhara et al., 2009). The 2012 NAEP 8th grade Writing Report Card (National Center for Educational Statistics, 2012) shows that only 25% of students' argumentative essays are at or above the proficient level. Empirical studies have revealed various problems when students were required to demonstrate their argumentation skills. For example, students often fail to include critical components (i.e., position, reason, evidence, counterarguments, and rebuttals) or to present them clearly (e.g., Ferretti et al., 2000). Other common problems include a lack of supporting evidence, a strong "my-side" bias, and poor essay organization (Felton & Kuhn, 2001; Ferretti et al., 2009; Kuhn et al., 1997).

The CCSS provide benchmarks for a variety of ELA skills that students are expected to master. As a result, teachers need valid and reliable measurements of students' argumentation skills that can help them identify student needs and determine appropriate instructional strategies and classroom practices. To elicit useful information about student skills for formative purposes, it is critical to ground assessments in disciplinary theories of learning (McNamara, 2011). Our project therefore aims to develop formative assessment tasks measuring students' argumentation skills that are grounded in disciplinary-based theories. We adopt an evidence-centered design process (ECD; Mislevy et al., 2003; Pellegrino et al., 2016; Zieky, 2014) to build explicit validity arguments that link test design decisions to inferences about student skills, informed by argumentation theories and practices.

We build an ELA literacy framework developed as part of the CBAL® research initiative (Cognitively Based Assessments *of*, *for* and *as* Learning; Bennett, 2010). Within this framework, argumentation is identified as a *key practice* – an integrated set of reading, writing, and critical thinking skills that must be coordinated to achieve specific literacy goals (Deane et al., 2015). Specifically, argumentation is conceptualized as the practice of *Discussing and Debating Ideas*, i.e., convincing an audience through logical argument (Deane & Song, 2015). Within the key

[1] CCRA.W.1; see http://www.corestandards.org/resources/key-points-in-english-language-arts.

practice framework, skills are not conceptualized in isolation, but as contributors to specific activities during specific phases of work, such as considering and evaluating multiple arguments, developing a thesis and building an argument to support it, or expressing one's argument in a specific written form. The key practice framework is designed to help teachers organize instruction around purpose-driven activities that integrate different skills, manage complex tasks using tailored instructional support, and identify targets for learning and assessment grounded in the learning sciences.

The key practices framework also postulates learning progression (LP) for specific skills, linked to the grade level expectations by standards such as the CCSS. Formative assessment tasks grounded in a disciplinary-based theory of LP within a domain can help teachers to make evidence-based inferences about students' skills and determine next steps in instruction (Bennett, 2011; Popham, 2008; Sparks & Deane, 2015). When LP for argumentation is linked with specific tasks, they may provide teachers with a useful framework for understanding student skill levels and determining appropriate next steps in instruction (Deane & Song, 2014, 2015). For example, if a student encountered difficulty in identifying other's opinions (as a level-1 skill in the LP), it means that this student lacks the foundational skill of understanding arguments and the teacher should provide explicit instruction, such as looking for stance markers in argumentative discourse.

In previous work undertaken as part of the CBAL initiative, multiple scenario-based assessments (SBAs) were developed to assess the ability to read, analyse, and produce written arguments (Deane et al., 2011; Deane & Song, 2014; van Rijn et al., 2014; Bennett et al., 2016). Each SBA was designed to simulate the process of reading source texts, creating and evaluating arguments, and writing an argument essay about a specific topic. Within the SBAs that we developed, we found that task difficulty was closely aligned with LP levels, and that each of the preparatory, or lead-in tasks contributed significantly to essay score prediction (Deane et al., 2019). However, due to the amount of reading and writing involved in the SBA design, only a small number of tasks could be included, focusing on a limited number of LP levels. Therefore, in this project, we aimed to cover a wider range of argumentation skills by developing and testing discrete tasks, on a variety of topics, aligned to specific LP levels.

We used the hypothesized LP in our key practices framework for *Discuss and Debate Ideas* (Deane & Song, 2015) to guide task development. We then conducted a study with several hundred middle school students to evaluate the measurement properties and performance of these LP-based tasks. We examined the degree to which the LP could be recovered empirically (i.e., how well patterns of student performance aligned with the theoretical LP levels specified in the key practices framework).

Literature review and theoretical framework

Development of argumentation skills

Argumentation is an inherently social activity involving dialogue between people who may hold different opinions about a controversial issue in order to achieve such pragmatic purposes as resolving a difference of opinions (pragma-dialectical approach; van Eemeren et al., 1996). By elementary- and middle-school age, students demonstrate considerable sensitivity to authority figures and can generate oral arguments that anticipate and address potential criticisms that could be leveled against them (Ferretti & Lewis, 2019). Writing studies also found that upper-elementary and middle school students include more argument elements if the writing task requires students to elaborate their goals for content and audience during planning and revision (e.g., Midgette et al., 2008).

Argumentation skill development depends critically on how well students can identify and select relevant evidence and explain how the evidence supports an argument (Brem & Rips, 2000; Kuhn et al., 1997). Many fifth or sixth graders can elaborate and provide details to support their arguments (Ferretti et al., 2000, 2009). For example, Kuhn and Crowell (2011) found that sixth graders become more aware of using relevant evidence to support their claims after training. However, some argumentation skills are challenging even for older students and may not develop before adulthood unless support is provided. For instance, even high school or college students may find it difficult to identify the assumptions behind people's arguments or integrate arguments from both sides of an issue (Klaczynski, 2000). Refuting opposing viewpoints and anticipating counterarguments is a difficult task even for adults, especially in a written context (Ferretti et al., 2000; Leitão, 2003; Nussbaum & Kardash, 2005). In a study with 472 students from three middle schools, Song et al., (2020) found that student performances on CBAL written critique tasks were generally low. Middle school students are at a critical, but still early stage of argumentation skill development. We are particularly interested in understanding where the middle schoolers are in that process: that is, what skills have they mastered, and where do they most need support?

The key practice: discuss and debate ideas

The CBAL key practice *Discuss and Debate Ideas* focuses on the argumentation skills required in academic reading and writing. In this key practice, students are expected to demonstrate skills (e.g., recognizing supporting information) and use strategies (e.g., asking questions to identify assumptions) to evaluate arguments from different perspectives and build arguments that support their position (Deane & Song, 2014). The process involves five distinguishing phases (types of activities that occur at different points in an extended argumentation process): understanding what is at stake in an issue, gathering relevant information, understanding different perspectives, developing and evaluating arguments, and then presenting arguments. Informed by a comprehensive review of research in the cognitive and learning sciences as well as the pragma-dialectical theory, each phase in our cognitive-developmental model focuses on different combinations of strategies and skills, as described below.

(1) *Understand the issue: context and stakes.* First, students should understand the context and the stakes of the issue to have a meaningful participation in argumentative discourse. They need to know who their audiences are and identify the audiences' interests and beliefs, so that they can select appropriate rhetorical strategies to persuade their audiences. The key skill in this phase is appeal building.

(2) *Explore the subject.* Students cannot make thoughtful arguments if they lack prior knowledge about the topic under discussion. They need to gather relevant information about the topic, which requires inquiry skills classified in another key practice – *Conduct Research and Inquiry* (see Sparks & Deane, 2015).

(3) *Consider positions*. People with sophisticated argumentation skills take into account different perspectives. However, students usually consider their own opinion and ignore what others think about the issue. Thoughtful consideration of alternative perspectives can help students determine which positions are reasonable and defensible. We use the term *taking a position* to identify this skill.

(4) *Create and evaluate arguments*. To defend a position, we must present strong reasons and relevant evidence to ensure that our arguments are logical or plausible. Thus, the critical skill in this phase is called *reasons and evidence*. This phase also involves the skills of evaluating other people's arguments to identify unwarranted assumptions that could undermine the logic or plausibility of their arguments, which will eventually strengthen our own arguments.

(5) *Organize and present arguments*. In this last phase, the critical skill focuses on presenting the arguments in an appropriate structure, informed by genre expectations and conventions, whether in informal conversation or in written text. We use the term *framing a case* to the skill of organizing and presenting arguments.

This key practice may help teachers gather evidence of student understanding and scaffold learning opportunities relevant to develop the skills and strategies in a particular phase. For example, under the phase of consider positions, if students only present their own position, the teacher could prompt them to consider alternative perspectives, and examine the issue from a different angle. For pedagogical purposes, a teacher might require students to undertake each of these five phases in sequence. However, the actual process of *Discuss and Debate Ideas* is flexible and recursive. Four of these phases in this model are specific to argumentation: appeal building, taking a position, reasons and evidence, and framing a case. We expect that the development in each of these skills will be strongly linked to the progress in the others, thus we perceive them as progress variables. Next, we describe how we developed the argumentation LP.

Argumentation learning progression

For the past several years, the CBAL team has sought to capture findings from argumentation literature and explicitly represent student argument skill development through an LP framework. We then test the LP by developing items intended to measure specific developmental levels (Bennett et al., 2016; Deane & Song, 2015; van Rijn et al., 2014). We built an argumentation LP that describes the types of evidence that would yield inferences of increasing sophistication of argumentation skills (Deane & Song, 2015). The argumentation LP helps establish the alignment with the subject-specific knowledge and skills described in the CCSS. For example, CCSS RI5.8 expects 5th graders to identify reasons and evidence that an author uses to support particular points in a text, while CCSS RI8.8 expects 8th-grade students to evaluate the soundness of reasoning and the relevancy and sufficiency of the evidence beyond simply identifying the argument components. Our argumentation LP also reflects such an advancement: at level 2, students can elaborate their reasons with some awareness of the need for evidence; at level 3, they can provide relevant evidence to support their points in a relatively logical way; when they move up to level 4, students can reason about and respond to counterarguments and critical questions. The LP is informed by the existing literature on argumentation skills as discussed above (e.g., reasons are developed before evidence, while counterarguments and rebuttals are more advanced), using available research evidence and prior literature about typical developmental trajectories for four argumentation skills identified in the framework (see Table 1 for a summary of the LP and its progress variables). The LP consists of five developmental levels (1: lowest, 5: highest). These five levels include progress variables that address three modes of cognitive processing: interpretive (reading), expressive (writing), and deliberative (critical thinking). Under each mode, specific descriptors of the target skills are provided to help inform the student, task, and evidence model components within an Evidence-Centered Design approach to assessment development (Mislevy et al., 2003). In summary, the LP serves multiple purposes in our research: as a description of major cognitive stages along a developmental continuum, as a framework for assessing where students are in argumentation skills, and as a sketch of how teachers might scaffold the development of argumentation skills step-by-step.

Although the LP is informed by existing studies, there are gaps in the literature, and other factors could influence the progression for individual students. Popham (2007) claimed that a universally accepted LP does not exist because individuals may follow different development trajectories toward mastery. Therefore, we expect to see individual differences in mastering various sub-skills (i.e., progress variables) rather than treating the LP as fixed developmental sequences. However, the LP generally reflects the relative sophistication of the skills, so most students will show evidence of mastery of lower-level skills before higher-level skills. Reaching a higher level typically means that the student gets to a higher level on multiple progress variables corresponding to the skills at each phase.

LPs are subject to empirical validation, challenge, and potential revision (Bennett, 2011; Corcoran et al., 2009). To validate the LPs, people need to develop

Table 1 Argumentation learning progression overview

	Appeal building	Taking a position	Reasons and evidence	Framing a case
Level 5	Displays a well-developed rhetorical (metacognitive) understanding of persuasion	Frames one’s own position in terms that exploit the current “state of discussion” (taking into account other positions)	Builds systematic mental models of entire debates, and uses that model to frame one’s own attempts to build knowledge	Displays mastery of different argument forms, demonstrating flexible control of genre features
Level 4	Develops rhetorical plans, with sensitivity to differences among audiences with different points of view	Recognizes unstated assumptions, biases, and other subjective elements in a text and develops one’s own position more clearly	Reasons about and responds to counterevidence and critical questions	Approaches persuasive text as part of a dialogue between multiple perspectives with attention to counterarguments
Level 3	Coordinates multiple appeals and moves into a coherent effort to persuade a target audience	Understands and expresses positions clearly, capturing their relationships both to similar and contrasting points of view	Provides evidence and reasons that are directly relevant to and support the main point in a logical way	Approaches persuasive text as a logically structured presentation of a case with embedded reasons and evidence
Level 2	Conducts simple analysis of how oneself or an author might appeal or has appealed to different audiences and interests	Understands and expresses positions with reasonable attention to what one knows and what is important in the domain	Elaborates on reasons with some awareness of the need for evidence	Approaches persuasive text as a coherently organized sequence of reasons supporting a position
Level 1	Tries to convince someone by making some sort of persuasive appeal	Shows a basic understanding of taking a side in an argument (Pro or Con)	Understands that positions need to be supported with reasons to convince the audience	Approaches argument as chain of individual turns, and understands and produces such turns in context

assessment tasks aligned to the LPs and collect empirical data to observe whether the performance patterns recover the general model and progress variables. In a recent study, Sparks et al., (2021) defined and tested their developmental trajectories for the source evaluation skills in the literacy practice, showing that LP tasks can yield useful information about students' literacy skill development. In this study, we intended to collect empirical evidence for the argumentation LP, and in what follows we provide a description of the LP-based assessment tasks design.

LP-based assessment tasks

To gather validity evidence for the argumentation LP, we designed a set of assessment tasks around the LP descriptions and progress variables and subjected those tasks to empirical evaluation. Designing assessment tasks aligned to the LP can support evidence-based inference about student achievement levels (Mislevy et al., 2003), which can be used to recommend classroom practices that scaffold students toward the next level of performance (Furtak, 2012). This approach has a potential to help solve the problems associated with low performance on traditional essay tests that provide relatively little information about why students fail to produce strong arguments (Hillocks, 2002). The assessments described in this paper are designed to provide information about students' strengths and weaknesses in argumentation.

The current study

A multidisciplinary team of research scientists and assessment specialists developed tasks that target the skills identified in the LP framework. The tasks included a range of items with the selected response (SR) format, the constructed response (CR) format, and a combination of SR and CR formats. The tasks were designed for computer delivery, including interactive item formats (e.g., drop-down menu, drag-and-drop, grid, matching, etc.), digital images, and glossary for unfamiliar words. Tasks included a variety of topics that assessment specialists judged as relevant and appropriate for middle schoolers. For example, the Appendix shows three sample LP-based tasks: School Newsletter is a level-1 task that asks students to identify supporting reasons for a given position, Protect Ears is a level-2 task that targets the skill of identifying supporting evidence in a text, and Monsters and Invaders is a level-3 task that assesses the skill of identifying and explaining logical flaws. Table 2 lists all the tasks by the LP level with a brief description, number of items, format, phase, and mode.

As LP levels increase, we expect to see increasing difficulty for students. However, processing modes may influence the task difficulty level. Expressive tasks can be more challenging than interpretive or deliberative tasks because they require more cognitive effort (McFarland et al., 1980). Even if targeting the same level skills, CR items that require students to generate written responses can be harder than SR items that merely require comprehension and interpretation of given options. The LP-based tasks went through content, editorial, and fairness reviews, conducted by

Table 2 Descriptions of LP-based tasks

LP level	Task name	Task description	# of items	Format	Phase	Mode	Test form
1	Pajama day	Identify appropriate argumentative discourse	7	SR	Frame	I	C
	Pizza party	Make simple appeals	2	SR	Appeal	D	B
	Movie review	Identify supporting reasons	2	SR	Reason	I	C
	Class newsletter	Develop supporting reasons	4	CR	Reason	D	A
	School newsletter	Identify supporting reasons	6	SR	Reason	I	A
	Music album	Identify appropriate argumentative discourse	8	SR	Frame	I	B
2	History field trip	Identify effective appeals	6	SR	Appeal	D	A
	Five-second rule	Identify supporting reasons and evidence	1	SR	Reason	I	C
	Social media	Infer the position and identify supporting points	2	SR	Position	I	C
	Protect ears	Identify supporting evidence	2	SR	Reason	I	C
	Tech effects	Identify effective appeals	6	SR	Appeal	D	C
	Empty building	Identify audience’s positions and values	3	SR&CR	Appeal	D	B
	School assembly	Identify audience’s positions and values	3	SR&CR	Appeal	D	A
	Space exploration	Infer the position and identify supporting points	3	SR	Position	I	A
	Voter turnout	Fill in an argument concept map	9	CR	Position	D	B
	Mystery novels	Organize arguments	4	SR	Frame	D	C
3	Cursive writing	Write a short argument paragraph	1	CR	Position	E	C
	Monsters V1	Identify and explain reasoning errors	3	SR&CR	Reason	D	A
	Monsters V2	Identify reasoning errors and supporting evidence	5	SR	Reason	D	B
	Foreign language	Write a short argument paragraph	1	CR	Position	E	B
	Mission to Mars	Organize arguments in an outline	3	SR	Frame	D	A
	Pluto	Identify the unsupported claim, and possible supports	2	SR	Reason	D	C
	Student orientation	Write appeals	1	CR	Appeal	E	C
	Governor visit	Write appeals	1	CR	Appeal	E	A
	Animal artworks	Identify the common ground and different opinions	9	SR	Position	D	B
	Curfew	Make an effective appeal to the audience	2	SR&CR	Appeal	D	C

Table 2 (continued)

LP level	Task name	Task description	# of items	Format	Phase	Mode	Test form
	Dinosaur fossil	Identify the unsupported claim, and possible supports	2	SR	Reason	D	B

I = Identification; E = Expression; D = Deliberation several experts. We also conducted a pilot study with 74 seventh-grade students and used the results to improve task design (e.g., clarifying task directions, correcting keyable distractors, and revising scoring rubrics).

The current study was aimed to gather preliminary validity evidence for the LP-based tasks. We evaluated how well these LP-based tasks function and explored the relationship between LP-based tasks and an external measure of students' ELA reading skills. We also examined what students' responses revealed about their argumentation skills, and how their performances might vary by grade level. Further, we addressed the empirical recovery of the argumentation LP for middle school students, by examining the order of the LP levels as they are assigned to the tasks. Our research questions included:

(1) Do the LP-based tasks provide reliable and valid measurement of students' argumentation skills?
(2) Are the patterns of student performance on the tasks aligned to the expectations by the intended LP levels?
(3) What factors in the LP-based item design could affect the task difficulty?

Method

Participants

Seven hundred and sixty students (grade 7: $N=433$; grade 8: $N=327$) from two suburban middle schools (School 1: $N=386$; School 2: $N=374$) in Western U.S. participated in the study. School-level demographic information was collected. The two schools have similar student demographic profiles. In School 1, roughly 45% of the students were Caucasian, 30% of the students received free or reduced-price lunch (a proxy for socioeconomic status), and 5% were English learners. In School 2, roughly 35% of the students were Caucasian; 31% of the students received free or reduced-price lunch; 6% were English learners. Further, both schools were far above the state average in the standardized tests in ELA and mathematics, suggesting that most students at these schools were performing at or above grade level.

Measures

Argumentation LP tasks

Twenty-seven tasks were organized into three test forms (A, B, C), covering all four phases of skills and three skill modes. Form A and Form B each contained eight tasks (18 and 25 items, respectively), and Form C had 11 tasks (18 items), across the first three LP levels. It took approximately 40 min to complete each form. The forms were randomly assigned to the students, with 659 students completing two forms and 101 students completing only one form. The forms had similar numbers of students (Form A: 472; Form B: 476; and Form C: 471).

Reading inventory and scholastic evaluation (RISE)[2]

The RISE was a computer-administered diagnostic reading assessment for students at grades 3–12. It was designed to help identify reading subskills students may lack so that teachers can provide additional instruction to them. The RISE contained six subtests, each of which targets a specific reading subskill, including word recognition and decoding, vocabulary, morphology, sentence processing, efficiency of basic reading comprehension, and reading comprehension (Sabatini et al., 2019). It took approximately one hour to complete the RISE, and student responses were scored by the computer.

Procedure

Students completed the RISE and argumentation LP-based tasks using school computers. LP-based tasks were presented online, with students randomly assigned to two forms. School staff were provided with directions for administering the tasks to students.

Scoring

SR items in LP-based forms were automatically scored, while CR items were human scored. Each CR item had a unique rubric developed to reflect the quality of the written response in terms of meeting the task requirements. Two experienced raters received the training to apply each scoring rubric. Training included review of rubrics, examination of benchmark responses, and practice with group scoring of sample responses. After the training, the raters scored responses independently, and the interrater agreement was measured by computing the kappa values. All but two of the CR items achieved an acceptable or good agreement ($k > = 0.70$). The average kappa was 0.83. Scores assigned by the first rater were used in the analysis. Two items in Voter Turnout had a low interrater agreement ($k = 0.46$, 0.53, respectively), so a third rater adjudicated the responses. We used the third rater's adjudicated scores in the analysis for these two items.

Data analysis

Data analysis first involved computing item-level descriptive statistics for each of the three test forms (score means and standard deviations, proportion correct, item-total correlations, Cronbach's alpha reliability). To assess concurrent validity, we computed correlations among test forms, and between each test form and the external measure of ELA reading skills (i.e., RISE). To address questions

[2] RISE is now called ReadBasix. https://www.captivoice.com/capti-site/public/entry/diagnostic

Table 3 LP form statistics

Form	N	Max. score	Obs. range	M	SD	$P+$	Cronbach's α
A	472	20.5	1.5–20.0	10.58	3.85	0.52	0.77
B	476	21.0	0.0–20.5	8.99	4.38	0.43	0.84
C	471	23.5	0.5–22.5	12.23	4.89	0.52	0.79

Table 4 Item analysis across LP forms

Form	p^+ range	Mean p^+	Median p^+
A	0.34 – 0.88	0.64	0.59
B	0.19 – 0.75	0.49	0.45
C	0.12 – 0.86	0.61	0.69

about empirical recovery of the LP, we employed an IRT model (the generalized partial credit model) to examine the LP level associated with each task. This approach involved task progression maps which are essentially generalizations of Wright maps (Wilson, 2011). These task progression maps have previously been shown to be effective for describing the recovery of LPs from assessment items in mathematics and ELA (see Deane & Song, 2014 for a case using ELA argumentation items).

Results

(RQ1) Do the LP-based tasks provide reliable and valid measurement of students' argumentation skills?

Reliability

Descriptive statistics for each test form are provided in Table 3. Overall Cronbach's alpha (internal consistency) reliability of Form B was good ($\alpha = 0.84$), and Cronbach's alphas of Form A and Form C were acceptable ($\alpha = 0.77$ and 0.79, respectively). Form B ($P^+ = 0.43$) appeared to be more difficult than the other two forms ($P^+ = 0.52$). The correlations among the forms were 0.72 (A&B), 0.70 (A&C), and 0.75 (B&C); after correcting for attenuation, those correlations were 0.89 (A&B), 0.90 (A&C), and 0.93 (B&C). These moderate to strong inter-correlations (all p's < .001) suggested that the forms measure similar constructs.

Further, item-level performance related to difficulty (p^+) were calculated as mean score divided by the maximum score of each item. Table 4 presents the range of p^+ values, and the mean and median p^+ values in each form. All three forms had a wide range of the item-level p^+ values (Form A: 0.34 to 0.88, Form B: 0.19 to 0.75, and Form C: 0.12 to 0.86), as we intentionally made it this way in our design. Two most challenging items were: Dinosaur Fossil ($p^+ = 0.19$) in

Table 5 Correlations between LP form performance (PCTN) and RISE score

Form	Grade	*N*	WRDC	VOC	MA	SEN	MZ	RC
A	All	472	0.51	0.59	0.56	0.55	0.64	0.67
	7	270	0.43	0.53	0.56	0.55	0.65	0.67
	8	202	0.58	0.58	0.49	0.51	0.57	0.64
B	All	476	0.52	0.62	0.59	0.54	0.62	0.67
	7	266	0.49	0.56	0.57	0.52	0.62	0.66
	8	210	0.53	0.65	0.59	0.56	0.60	0.66
C	All	471	0.53	0.55	0.58	0.59	0.64	0.70
	7	263	0.49	0.50	0.57	0.57	0.63	0.70
	8	208	0.58	0.59	0.57	0.60	0.63	0.69

WRDC = word recognition and decoding, VOC = vocabulary, MA = morphology, SEN = sentence processing, MZ = (MAZE) efficiency of basic reading, RC = reading comprehension

Form B, and Curfew ($p^+ = 0.12$) in Form C. Both items consisted of two questions and the answer to the second question was contingent to the answer to the first question (i.e., if a student answered the first question incorrectly, he would also answer the second question incorrectly). The easiest items included: four items in the History Field Trip task ($p^+ = 0.87$ or 0.88) in Form A, and Five-Second Rule ($p^+ = 0.86$) in Form C.

Finally, we calculated biserial (or polyserial) correlations between item and corrected total score for each item. The Pizza Party second item in Form B had the lowest item-total correlation (0.11), while all others exceeded 0.20.

Concurrent validity

Correlations were computed between students' performance on each LP form and their RISE Scores (see Table 5). Correlations ranged from 0.51 to 0.70 (all p's < .001). Student performance on the LP forms appeared to have highest correlations with reading comprehension scores in the RISE assessments (0.67, 0.67, 0.70, for the three forms, respectively), and relatively moderate correlations with the measure of efficiency of basic reading comprehension (all above 0.60). Correlations by grade showed the similar trend. However, 8th graders' LP task performance appeared to have a slightly higher correlation with word recognition and decoding as well as vocabulary knowledge, compared to 7th graders.

(RQ2) Are the patterns of student performance on the tasks aligned to the expectations by the intended LP levels?

To recover the LP, we conducted an IRT analysis that placed and compared the tasks on a common scale and provided estimates of item difficulty and student ability. A generalized partial credit model (GPCM; Muraki, 1992) was fitted to the data

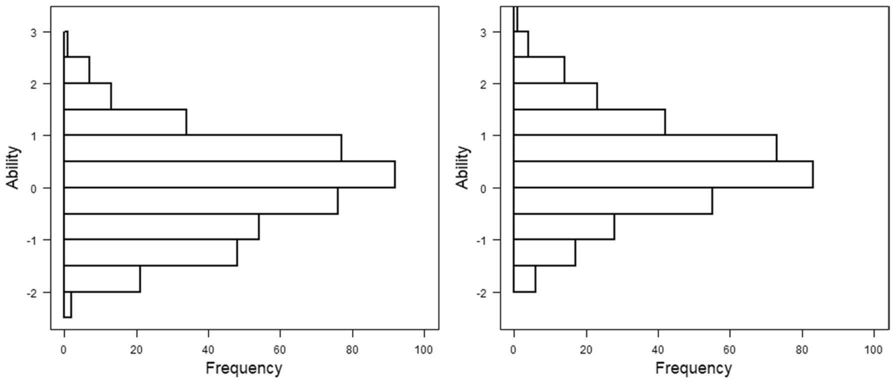


Fig. 1 Histogram of student ability estimates derived from the GPCM model (Left panel: seventh grade, Right panel: eighth grade).

Table 6 Performance (*P*+and mean scores) by grade

Form	MAX score	Grade	*N*	Mean score	*SD*	*P*+
A	20.5	7	270	9.65	3.69	0.47
		8	202	11.81	3.70	0.58
B	21.0	7	266	8.19	4.07	0.39
		8	210	10.00	4.53	0.48
C	23.5	7	263	11.58	4.92	0.49
		8	208	13.06	4.72	0.56

using open-source MIRT software developed by Haberman (2013)[3]. The GPCM is appropriate for modeling items with partial credit, which many of our LP tasks included. The model includes an overall discrimination parameter for each item and item-category-specific intercept parameters for each non-zero score. A predictor for grade level was included in order to differentiate between seventh- and eighth-grade students.

We also ran a multidimensional GPCM using the four progress variables as dimensions and found that the correlations among the four LP variables were very high (ranging from 0.86 to 0.98), which confirmed that the progress variables represent the same construct. We therefore chose a unidimensional model to illustrate further results. We obtained two sets of estimates from the GPCM analysis: student ability estimates (expected a postiori, or EAP estimates), and task progression maps based on estimated item parameters.

Student ability estimates reflect the estimated range of ability levels of the participating students based on their task performance. Figure 1 presents a

[3] The program, manual, examples, and source code are freely available at https://github.com/EducationalTestingService/MIRT.

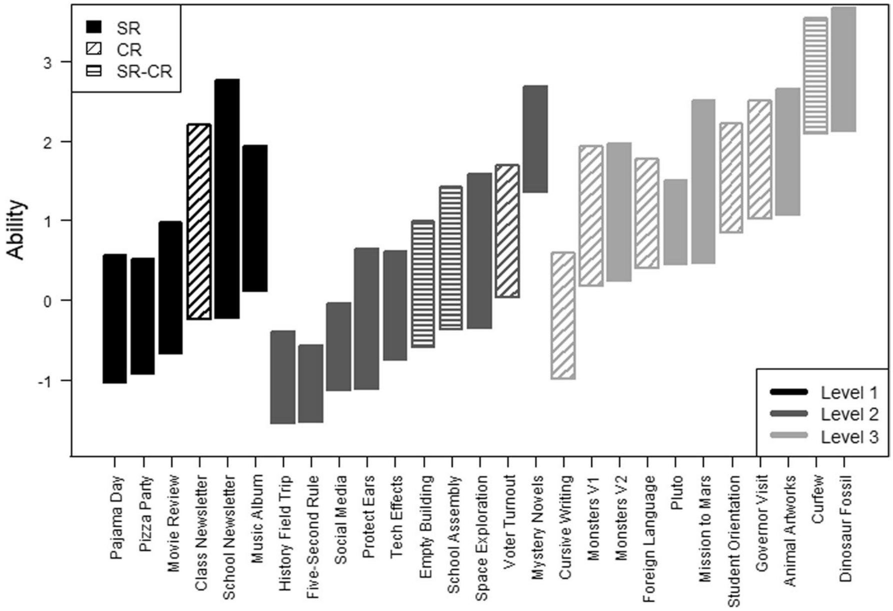


Fig. 2 Task progression map. Each rectangle indicates the student ability interval that is linked to a 50-80% expected task score under the model

histogram of the student ability estimates derived from the GPCM model. The left panel shows the histogram for seventh-grade students, and the right panel shows the histogram for eighth-grade students. As it shows, the student ability estimates roughly ranged from -3 to $+3$. In the GPCM analysis, eighth-grade students outperformed seventh-grade students (Cohen's $d = .45$). For each form, eighth graders had a higher mean score than seventh graders, and the difference was statistically significant (with a small to medium effect size), indicated by the independent-samples t-tests [Form A: $t(470) = -6.28$, $p < .001$; *Cohen's* $d = .58$; Form B: $t(474) = -4.53$, $p < .001$; *Cohen's* $d = .42$; Form C: $t(469) = -3.30$, $p < .010$; *Cohen's* $d = .30$]. See Table 6 for form-level statistics by grade.

Then we created a task progression map to illustrate alignment among estimates of task difficulty and student ability, taking the LP level and item format into account. Figure 2 shows the task progression map that included 27 tasks under the unidimensional GPCM. Each rectangle indicates the student ability interval that is linked to a 50–80% expected task score under the model (van Rijn et al., 2014). The task progression map ranged from slightly above -2 to slightly above $+3$, mostly overlapping with the student ability range. As Fig. 2 shows, the intended ordering of the LP levels did not always align well with the ordering of the task progression maps. Sometimes, certain higher-level tasks were found to be easier than lower-level tasks (e.g., History Field Trip, Five Second Rule, Teaching Cursive). Level-3 tasks generally appeared to be more difficult than level-1 and level-2 tasks, but the difficulty of level-1 and level-2 tasks were quite similar. We had 17 SR tasks, seven CR tasks, and three tasks that used the combination of the two formats. All three formats had a wide range of difficulty. While there was no clear pattern for which format was more challenging, the easiest tasks appeared to be the SR format.

(RQ3) What factors in the LP-based item design could affect the task difficulty?

Most of the tasks performed well and were discriminating. In this section we focus on items with notable issues. The second item in Pizza Party was the only item that did not have adequate item-total correlation ($r = .11$). This task sets up a brief context [your class is going to have a party, and you need to convince your teacher that it will be easy to have a pizza party (item 1) and convince your classmates that it will be fun to have a pizza party (item 2)]. The task assesses whether students can differentiate appeals by their likely effectiveness for different audiences in a familiar context. Students were given five reasons and asked to select the most convincing reason for each item. Roughly two-thirds of the students successfully selected the most convincing reason to the teacher, while only 42% of the students identified the most convincing reason for their classmates (i.e., our class thinks pizza is delicious, and we would love being able to choose our own topping). Ninety-four students (20%) selected the first option (i.e., my cousins love pizza; we eat it almost every time we go to their house). This option was a distractor: even though it also supported that the pizza party is fun, it was not the best answer because it did not directly address to the intended audience. However, some students might perceive this connecting to their personal experience.

Two tasks were found to be very challenging: Dinosaur Fossil ($p^+ = 0.19$) in Form B, and Curfew ($p^+ = 0.12$) in Form C. Both are level-3 tasks. Dinosaur Fossil targets the Reason and Evidence skill. Students first read a short passage to identify an unsupported claim. Then from a given list of information, students need to identify evidence that could support the claim. Curfew targets the Appeal Building skill. Students are presented with a brief scenario that shows the city council is considering a curfew. Students are told that they are opposed to the curfew and need to select the argument that will most likely convince the council members. Then they need to explain why that argument will be the most appealing to the intended audience. Both tasks consist of two questions, and if students could not answer the first question correctly, they would automatically receive a 0 on the task. A significant percentage of students did not answer the first question correctly, and therefore received no credit for this task.

Our IRT analysis showed that the student ability range and the task difficulty range were mostly well aligned. Level-3 tasks in general were more difficult than lower-level tasks except for Teaching Cursive. However, level-1 and level-2 tasks had similar difficulties. There are at least two possible explanations for this. First, we had fewer level-1 tasks in the forms because middle school students are expected to have mastered level 1. Most level-1 tasks in our task pool were developed for younger students, so the topic and text were not suitable for middle school students. Therefore, we only included tasks suitable for middle school level, and these tasks might be at the higher end of level 1 and close to level 2. Second, the scoring rules may have contributed inappropriately to task difficulty. For example, the Pajama Day task (level 1) consisted of seven statements, and for each statement, student need to make a judgment about whether the statement was helpful or not. The scoring rule was: students receive 2 points if they made correct judgments on all seven statements, 1 point if they made correct judgments on six statements, and 0 if they made 5 or fewer correct judgments. Based on this scoring rule, the $p+$ value for the Pajama Day task was 0.72. However, with a closer examination, we found that only one item was somewhat difficult ($p^+ = 0.72$), and the remaining six items were very easy ($p^+ > 0.90$). Therefore, the current scoring method might overestimate the difficulty of this task.

Three tasks appeared to be easier than other tasks at their assigned level: History Field Trip and Five Second Rule at level 2, and Cursive Writing at level 3. Interestingly, these tasks were designed from the same or similar blueprints as Technology Effect, Protect Ears, and Foreign Language, respectively. Those tasks appeared to be harder, especially the Foreign Language task. We suspect that the topics contributed to the difficulty of the tasks. For example, in Cursive Writing and Foreign Language, students read a short background article about the given issue. Then they wrote a paragraph that includes their position on the issue with at least two reasons and a consideration of the opposing position. The design of the two tasks was parallel, which makes the choice of topic obvious differentiating variable. Students might be relatively familiar with cursive writing as a topic, but have very little knowledge about the relationship between foreign language learning and age.

Discussion

In this project, we collected empirical evidence from several hundred middle school students to evaluate a set of LP-based items designed to assess ELA argumentation skills. Our analyses focused on item reliability and validity, student performance patterns (grade level, RISE measures, LP level), and task alignment to LP levels. We also analysed factors that could have contributed to task difficulty.

Our reliability and concurrent validity analyses revealed that the assessment tasks were reliable, with moderate to high correlations observed among forms, adequate item-total correlations, and adequate correlations with external measures of reading subskills. These results provided preliminary evidence that the LP-based tasks are related to students' general ELA reading skills. Students' LP performance scores had the strongest correlation with RISE reading comprehension scores, which makes sense, since both argumentation and reading comprehension are higher-order literacy skills. Argumentation also appeared to presuppose adequate performance on the RISE subtests for foundational reading skills, such as word recognition, decoding, and vocabulary. This is not surprising, as, for example, Wang and his colleagues (2019) found that students below a decoding threshold in the RISE test did poorly on reading comprehension and showed little progress in reading comprehension in the following years. Therefore, teachers need to provide support to struggling readers to develop foundational reading skills so that they have the capability to grow in the argumentation space. Engaging struggling readers in an argumentative discourse might also be a solution. Teachers may have students orally discuss a controversial issue to facilitate their understanding of the given issue and the development of students' arguments.

As we expected, the eighth graders outperformed seventh graders on all three assessment forms. The consistently higher performance among eighth graders is also aligned to our LP theory, as students are expected to develop more sophisticated argumentation skills over time. However, the results only partially confirmed the items' hypothesized LP levels. Even though level-3 tasks were generally more challenging than level-1 and level-2 tasks, our level-1 and level-2 tasks had similar difficulty ranges, and some tasks were easier or harder than expected. Multiple factors may have contributed to this variability in the difficulty of items at the same LP level, including task design, scoring rule, and topic. In terms of design, tasks with two or more parts could be more challenging, especially when a follow-up question connects to a previous question. For example, if a student fails to identify an unsupported claim, he probably would not identify the evidence that helps support the right claim. Scores may underestimate student ability when the scoring rule requires a correct SR response to earn credit for a subsequent CR response (Sparks et al., 2021). In addition, students' prior knowledge about the topic could affect their performance. Even though we used parallel designs for some tasks, student performance still varied. Knowledge about the topic plays a substantial role in writing performance (e.g., Bereiter & Scardamalia, 1987; McCutchen et al., 1997), and impacts students' reading comprehension (McCarthy et al., 2018; McCrudden et al., 2016; O'Reilly & Sabatini, 2013). Therefore, lack of topic knowledge can lead to ineffective argumentation in students' written responses. We plan to refine the task design and scoring methods based on the data to minimize the impact of unexpected factors contributing to task difficulty, and thus achieve a better alignment with the argumentation LP framework. Since much the item variability related to the topics and students' prior knowledge and interest in the topics, we believe we can replicate our success in our prior studies using scenario-based designs (Bennett et al., 2016; Deane et al., 2019). Combining items designed to measure different LP levels in a single scenario will allow students to work through a series of tasks with varying difficulty levels on the same topic. In future work, we will analyse the sources of item difficulty in greater depth to identify whether revision of the argumentation LP is necessary.

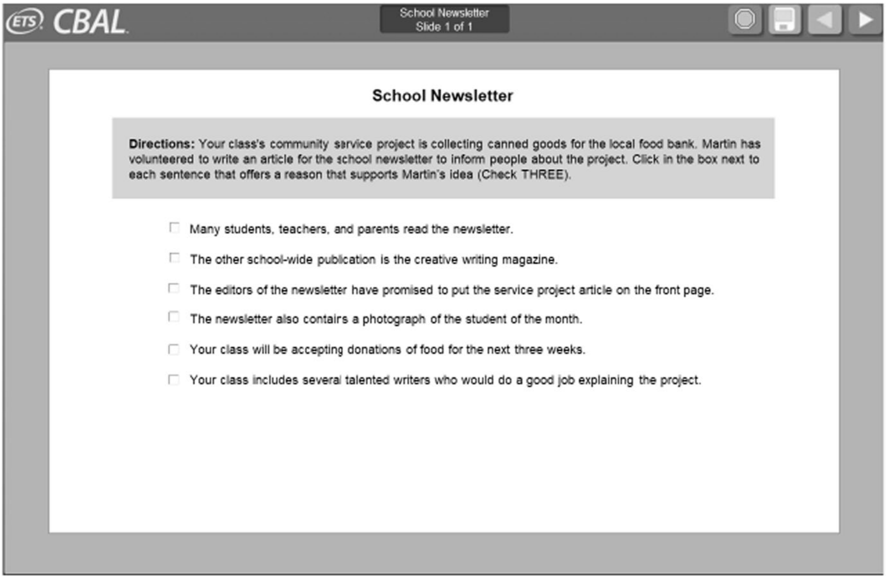
One limitation of this study relates to the samples, which were drawn from two suburban schools in the Western U.S. These samples might not be representative of schools in other areas. For example, student ability estimates might vary if we collect data in schools that have a large proportion of minorities or English learners. Another limitation is that the background data was obtained at the school level. If we had been able to collect individual student background information and prior year grades or test scores, we would have been able to conduct additional analyses, such as subgroup performance comparisons and examinations of correlations between LP-based task performance and academic achievement. We did collect RISE scores, but the RISE assessment only measures foundational reading skills. Due to time constraints on task administration, we were not able to include LP-based tasks that required students to compose a complete essay. The only CR tasks were able to include were short constructed-response items. As a result, we did not provide independent external measures of writing skills, which might also have affected student performance on the LP-based tasks. In future investigations, we will either implement an independent assessment of writing or collect participant's writing performance in a standardized assessment.

Despite these limitations, we gathered preliminary evidence for the LP approach. The current study was focused on detecting the overall progression of the argument construct, and our data suggested that the specific argumentation skills generally developed at a similar pace. In this study, we examined the variability of items at the same LP level across tasks and topics and found there is a considerable variation. However, in earlier studies we found items at different LP levels in the same scenario showed a progression of difficulty (e.g., Deane et al., 2019). This suggests much of the variability might be due to the scenario/topic or to extrinsic sources of variation such as task format or scoring method. We intend to explore this hypothesis in future research. If this hypothesis is correct, it should be possible to use LP levels to inform instructors where students may need support or instruction to be able to move up to the next level (Bennett et al., 2016; Deane et al., 2019).

The LP-based tasks we developed in this study primarily measure argument-related reading skills. These skills are important to support source-based argumentation. Going forward, we will explore ways to align measurement of LP level with performance on writing tasks, following up on results on our prior studies, but leveraging the variety of item types we developed for this study. It will be particularly important to conduct classroom observations to understand how teachers can use information provided by LP-based tasks to determine skill levels and decide next steps in instruction. Results from such work will have important implications for the way that LP can be used to interpret student learning achievements and determine next goals for instruction, as a theoretically grounded formative assessment process (Sparks et al., 2021).

Appendix: Sample LP-based tasks

School newsletter (Level 1)



School Newsletter

Directions: Your class's community service project is collecting canned goods for the local food bank. Martin has volunteered to write an article for the school newsletter to inform people about the project. Click in the box next to each sentence that offers a reason that supports Martin's idea (Check THREE).

□ Many students, teachers, and parents read the newsletter.

□ The other school-wide publication is the creative writing magazine.

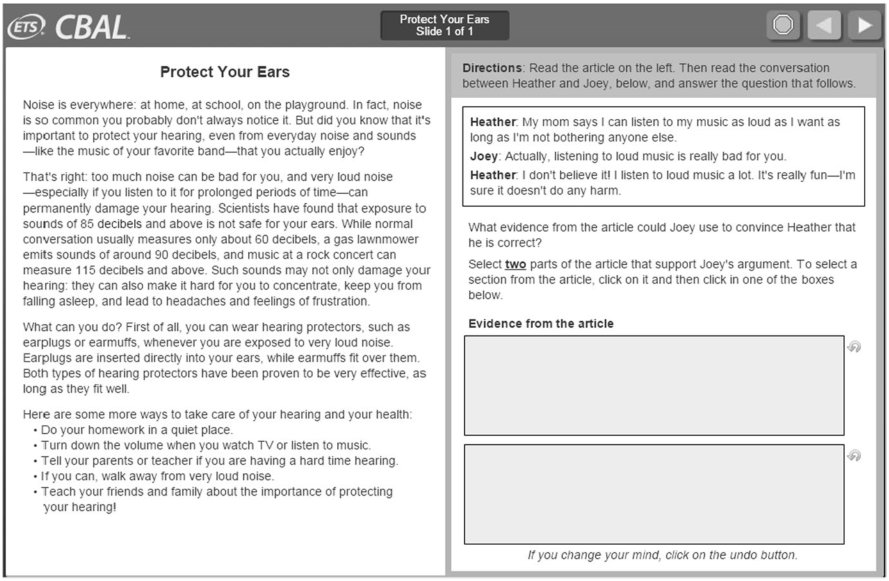
□ The editors of the newsletter have promised to put the service project article on the front page.

□ The newsletter also contains a photograph of the student of the month.

□ Your class will be accepting donations of food for the next three weeks.

□ Your class includes several talented writers who would do a good job explaining the project.

Protect ears (Level 2)



Protect Your Ears

Noise is everywhere: at home, at school, on the playground. In fact, noise is so common you probably don't always notice it. But did you know that it's important to protect your hearing, even from everyday noise and sounds —like the music of your favorite band—that you actually enjoy?

That's right: too much noise can be bad for you, and very loud noise —especially if you listen to it for prolonged periods of time—can permanently damage your hearing. Scientists have found that exposure to sounds of 85 decibels and above is not safe for your ears. While normal conversation usually measures only about 60 decibels, a gas lawnmower emits sounds of around 90 decibels, and music at a rock concert can measure 115 decibels and above. Such sounds may not only damage your hearing: they can also make it hard for you to concentrate, keep you from falling asleep, and lead to headaches and feelings of frustration.

What can you do? First of all, you can wear hearing protectors, such as earplugs or earmuffs, whenever you are exposed to very loud noise. Earplugs are inserted directly into your ears, while earmuffs fit over them. Both types of hearing protectors have been proven to be very effective, as long as they fit well.

Here are some more ways to take care of your hearing and your health:

- Do your homework in a quiet place.
- Turn down the volume when you watch TV or listen to music.
- Tell your parents or teacher if you are having a hard time hearing.
- If you can, walk away from very loud noise.
- Teach your friends and family about the importance of protecting your hearing!

Directions: Read the article on the left. Then read the conversation between Heather and Joey, below, and answer the question that follows.

Heather: My mom says I can listen to my music as loud as I want as long as I'm not bothering anyone else.

Joey: Actually, listening to loud music is really bad for you.

Heather: I don't believe it! I listen to loud music a lot. It's really fun—I'm sure it doesn't do any harm.

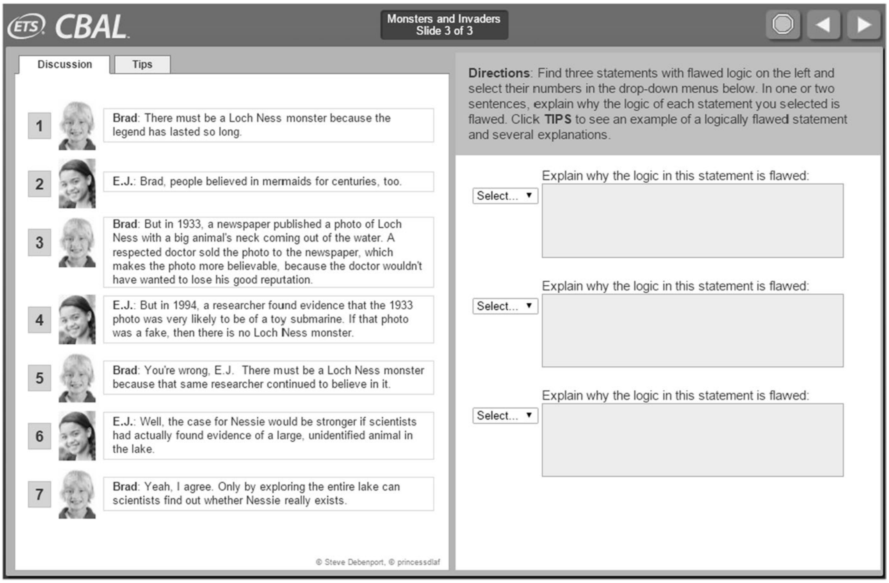
What evidence from the article could Joey use to convince Heather that he is correct?

Select **two** parts of the article that support Joey's argument. To select a section from the article, click on it and then click in one of the boxes below.

Evidence from the article

If you change your mind, click on the undo button.

Monsters and invaders (Level 3)



1 **Brad**: There must be a Loch Ness monster because the legend has lasted so long.

2 **E.J.**: Brad, people believed in mermaids for centuries, too.

3 **Brad**: But in 1933, a newspaper published a photo of Loch Ness with a big animal's neck coming out of the water. A respected doctor sold the photo to the newspaper, which makes the photo more believable, because the doctor wouldn't have wanted to lose his good reputation.

4 **E.J.**: But in 1994, a researcher found evidence that the 1933 photo was very likely to be of a toy submarine. If that photo was a fake, then there is no Loch Ness monster.

5 **Brad**: You're wrong, E.J. There must be a Loch Ness monster because that same researcher continued to believe in it.

6 **E.J.**: Well, the case for Nessie would be stronger if scientists had actually found evidence of a large, unidentified animal in the lake.

7 **Brad**: Yeah, I agree. Only by exploring the entire lake can scientists find out whether Nessie really exists.

© Steve Debenport, © princessdlaf

Directions: Find three statements with flawed logic on the left and select their numbers in the drop-down menus below. In one or two sentences, explain why the logic of each statement you selected is flawed. Click **TIPS** to see an example of a logically flawed statement and several explanations.

Select... ▾ Explain why the logic in this statement is flawed:

Select... ▾ Explain why the logic in this statement is flawed:

Select... ▾ Explain why the logic in this statement is flawed:

Acknowledgements We gratefully acknowledge funding from the ETS CBAL® research initiative. We would like to thank Kelly Bruce for administrating the LP forms with the participating schools. We also want to thank Tenaha O'Reilly and Heather Buzick for providing feedback on earlier versions of this manuscript.

Funding This work was supported by Educational Testing Service under the Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL®) Research Initiative.

Declaration

Conflict of interest The authors declare that they have no conflict of interest..

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*, 70–91. https://doi.org/10.1080/15366367.2010.508686.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678.

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.

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

Brem, S., & Rips, L. (2000). Evidence and explanation in informal argument. *Cognitive Science*, *24*(4), 573–604. https://doi.org/10.1207/s15516709cog2404_2.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* New York, NY: Center on Continuous Instructional Improvement, Teachers College—Columbia University. https://doi.org/10.12698/cpre.2009.rr63

Council of Chief State School Officers & National Governors Association (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects* Washington, DC: Author. Retrieved from www.corestandards.org/the-standards/ELA-Literacy

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

Deane, P., Sabatini, J., Feng, G., Sparks, J. R., Song, Y., Fowles, M., O'Reilly, T., Jueds, K., Krovetz, R., & Appel, C. (2015). *Key practices in the English language arts (ELA): Linking learning theory, assessment, and instruction.* (Research Report RR-15-17). Princeton, NJ: Educational Testing Service. 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. *Spanish Journal of Educational Psychology (Psicologia Educativa), 20*, 99−108.

Deane, P., & Song, Y. (2015). *The key practice, 'discuss and debate ideas': Conceptual framework, literature review, and provisional learning progressions for argumentation.* (Research Report RR-15-33). Princeton, NJ: Educational Testing Service.

Deane, P., Song, Y., van Rijn, P., O'Reilly, T., Bennett, R. E., Fowles, M., Sabatini, J., & Zhang, M. (2019). The case for scenario-based assessment of written argumentation. *Reading and Writing, 32*, 1575–1606.

Felton, M., & Kuhn, D. (2001). The development of argumentive discourse skills. *Discourse Processes*, *32*(2–3), 135–153. https://doi.org/10.1207/S15326950DP3202&3_03.

Ferretti, R. P., & De La Paz, S. (2011). On the comprehension and production of written texts: instructional activities that support content-area literacy. In R. O'Connor, & P. Vadasy (Eds.), *Handbook of reading interventions* (pp. 326–355). New York, NY: Guilford.

Ferretti, R. P., & Lewis, W. E. (2019). Knowledge of persuasion and writing goals predict the quality of children's persuasive writing. *Reading and Writing*, *32*(6), 1411–1430. https://doi.org/10.1007/s11145-018-9918-6.

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.

Ferretti, R. P., MacArthur, C. A., & Dowdy, N. S. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology*, *92*(4), 694–702. https://doi.org/10.1037/0022-0663.92.4.694.

Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching*, *49*(9), 1181–1210. https://doi.org/10.1002/tea.21054.

Gilbert, J., & Graham, S. (2010). Teaching writing to elementary students in grades 4–6: a national survey. *Elementary School Journal*, *110*, 494–518. https://doi.org/10.1086/651193.

Gleason, M. M. (1999). The role of evidence in argumentative writing. *Reading & Writing Quarterly*, *15*(1), 81–106. https://doi.org/10.1080/105735699278305.

Gleason, M. M., & Isaacson, S. (2001). Using the new basals to teach the writing process: modification for students with learning problems. *Reading & Writing Quarterly*, *17*(1), 75–92. https://doi.org/10.1080/105735601455747.

Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M. A., Greenleaf, C., Lee, C. D., Shanahan, C., & Project, R. E. A. D. I. (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.

Graff, G. (2003). *Clueless in academe: how schooling obscures the life of the mind*. New Haven, CT: Yale University Press.

Graham, S., & Harris, K. R. (2013). Common core state standards, writing, and students with LD: recommendations. *Learning Disabilities Research & Practice*, *28*(1), 28–37. https://doi.org/10.1111/ldrp.12004.

Haberman, S. J. (2013). A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm (ETS RR-13-32). Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-13-32.pdf

Kiuhara, S., Graham, S., & Hawken, L. (2009). Teaching writing to high school students: a national survey. *Journal of Educational Psychology*, *101*(1), 136–160. https://doi.org/10.1037/a0013097.

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. https://doi.org/10.1111/1467-8624.00232.

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 argumentive reasoning. *Cognition and Instruction*, *15*(3), 287–315. https://doi.org/10.1207/s1532690xci1503_1.

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.

McCarthy, K. S., Guerrero, T. A., Kent, K. M., Allen, L. K., McNamara, D. S., Chao, S. F., Steinberg, J., O'Reilly, T., & Sabatini, J. (2018). Comprehension in a scenario-based assessment: domain and topic specific background knowledge. *Discourse Processes*, *55*(5–6), 510–524. https://doi.org/10.1080/0163853X.2018.1460159.

McCrudden, M. T., Stenseth, T., Bråten, I., & Strømsø, H. I. (2016). The effects of topic familiarity, author expertise, and content relevance on norwegian students' document selection: a mixed methods study. *Journal of Educational Psychology*, *108*(2), 147–162.

McCutchen, D., Francis, M., & Kerr, S. (1997). Revising for meaning: Effects of knowledge and strategy. *Journal of Educational Psychology*, *89*(4), 667–676. https://doi.org/10.1037/0022-0663.89.4.667.

McFarland, C. E., Frey, T. J., & Rhodes, D. D. (1980). Retrieval of internally versus externally generated words in episodic memory. *Journal of Verbal Learning & Verbal Behavior*, *19*(2), 210–225. https://doi.org/10.1016/S0022-5371(80)90182-6.

McNamara, T. (2011). Applied linguistics and measurement: a dialogue. *Language Testing*, *28*(4), 435–440. https://doi.org/10.1177/0265532211413446.

Midgette, E., Haria, P., & MacArthur, C. (2008). The effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eighth-grade students. *Reading and Writing*, *21*(1–2), 131–151. https://doi.org/10.1007/s11145-007-9067-9.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–67. https://doi.org/10.1207/S15366359MEA0101_02.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206.

National Center for Education Statistics. (2012). *The Nation's Report Card: writing 2011*. Washington, D.C.: Institute for Education Sciences, U.S. Department of Education. (NCES 2012 – 470.

Nussbaum, M. E., & Kardash, C. M. (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., & Sabatini, J. (2013). *Reading for understanding: How performance moderators and scenarios impact assessment design* (Research Report RR-13-31). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02338.x

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550.

Popham, W. J. (April 2007). The lowdown on learning progressions. *Educational Leadership*, *64*(7), 83–84.

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: ASCD.

Sabatini, J., Weeks, J., O'Reilly, T., Bruce, K., Steinberg, J., & Chao, S. (2019). *SARA reading components tests, RISE forms: Technical adequacy and test design (3rd Edition)* Research Report RR-19-36. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12269

Samudra, P. G. (2017). How can I persuade you? The development of audience awareness in children's oral and written arguments. [Doctoral Dissertation, University of Michigan]. Retrieved from https://deepblue.lib.umich.edu/bitstream/handle/2027.42/138639/preetigs_1.pdf?sequence=1

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: rethinking content-area literacy. *Harvard Educational Review*, *78*(1), 40–59.

Song, Y., Chao, Szu-Fu, & Attali, Y. (2020). Exploring the effect of a scaffolding design on students' argument critique skills. *Informal Logic, 40*(4), 605–628.

Sparks, J. R., & Deane, P. (2015). *Cognitively based assessment of research and inquiry skills: Defining a key practice in English language arts.* (Research Report RR-15-35). Princeton, NJ: Educational Testing Service.

Sparks, J. R., van Rijn, P., & Deane, P. (2021). Assessing source evaluation skills of middle school students using learning progressions. *Educational Assessment, 26*(4), 213–240. https://doi.org/10.1080/10627197.2021.1966299

van Eemeren, F. H., Grootendorst, R., & Henkemans, F. S. (1996). *Fundamentals of argumentation theory: a handbook of historical backgrounds and contemporary developments*. Mahwah, NJ: Erlbaum.

van Rijn, P. W., Graf, E. A., & Deane, P. (2014). Empirical recovery of argumentation learning progressions in scenario-based assessments of English language arts. *Spanish Journal of Educational Psychology (Psicologia Educativa), 20*, 109−115.

Wang, Z., Sabatini, J., O'Reilly, T., & Weeks, J. (2019). Decoding and reading comprehension: a test of the decoding threshold hypothesis. *Journal of Educational Psychology*, *111*(3), 387–401. https://doi.org/10.1037/edu0000302.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Spanish Journal of Educational Psychology (Psicologia Educativa)*, *20*(2), 79–87. https://doi.org/10.1016/j.pse.2014.11.003.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.