Designing Accessible Formative Assessment Tasks to Measure Argumentation Skills for English Learners

ETS RR–19-15

Danielle Guzman-Orth
Yi Song
Jesse R. Sparks

December 2019

ETS Research Report Series

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

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

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Designing Accessible Formative Assessment Tasks to Measure Argumentation Skills for English Learners

Danielle Guzman-Orth,[1] Yi Song,[2] & Jesse R. Sparks[2]

1 Educational Testing Service, Sacramento, CA
2 Educational Testing Service, Princeton, NJ

In this study, we investigated the challenges and opportunities in developing a computer-delivered English language arts (ELA) task intended to improve the accessibility of the task for middle school English learners (ELs). Data from cognitive labs with 8 ELs with varying language proficiency levels provided rich insight to student–task interaction and how the accessibility of the task could be improved to enhance student understanding and to support valid integration of the task as a part of students' formative assessment process. In this paper, we share the results from our research and discuss our iterative approach to improving ELA task accessibility for ELs' formative assessment process.

Keywords English learners; formative assessment; accessibility; argumentation; computer-delivered tasks; scenario-based assessment; English language arts

doi:10.1002/ets2.12251

Formative assessment is gaining popularity as a method to provide accurate, timely, and actionable information that can be used by teachers and students to improve learning (Heritage, 2010; Heritage, Kim, Vendlinski, & Herman, 2009). For English learners (ELs), this process may be confounded by their varying levels of English language proficiency (ELP), and teachers may be challenged by the need to provide appropriate supports to meet the diverse needs of their students (Shore, Wolf, & Blood, 2013). Informed by learning progressions (LPs) and the Common Core State Standards (Council of Chief State School Officers [CCSSO] & National Governors Association [NGA], 2010), we designed this study to investigate how ELs interact with a scenario-based assessment of English language arts (ELA) argumentation skills as part of their formative assessment process. This paper presents our theoretical framework and an overview of our multipronged research process, including a discussion of findings from our cognitive lab investigation and literature review. We also provide a detailed summary of the design process for the revised activity that focuses on the same argument construct while increasing overall accessibility.

Theoretical Framework

Argumentation skills are essential for success in college, career, and life; therefore, these skills have played a prominent role in recent educational reform efforts such as the Common Core State Standards (CCSSO & NGA, 2010). Despite the importance of argumentation skills, many students cannot write sound arguments or critically evaluate arguments, as evidenced by a variety of large-scale assessments and empirical studies (e.g., Ferretti, Lewis, & Andrews-Weckerly, 2009; National Center for Education Statistics, 2012; Perkins, Farady, & Bushey, 1991; Song, Deane, & Fowles, 2017). One of the most complex academic skills, argumentation has not been well supported in instructional practice, which frequently emphasizes basic written composition and use of specific templates while doing little to develop arguments and critical thinking (Hillocks, 2002). Furthermore, traditional assessments of argumentation, which typically require students to write an essay on a single prompt, offer little information about why students may have failed to accomplish this task.

To gather relevant evidence about students' argumentation skills, we designed a scenario-based assessment aligned to a set of hypothesized LPs (Deane & Song, 2015), as part of the *CBAL*® research initiative (cf. Bennett, 2010). Informed by cognitive and learning sciences research (e.g., Bereiter & Scardamalia, 1987; Graham & Perin, 2007; Hayes & Flower, 1980; Kuhn, 1991), argumentation LPs describe how argumentation skills develop into sophistication, characterizing the

Corresponding author: D. Guzman-Orth, E-mail: dguzman-orth@ets.org

qualitative shifts that occur as students reach higher levels in four strands of skills: (a) *appeal building:* understanding an audience's values and beliefs; (b) *taking a position:* developing a position and understanding other perspectives; (c) *reasons and evidence:* using reasons and evidence to support an argument and to evaluate others' arguments; and (d) *framing a case:* organizing and presenting an argument logically. We used LPs as a general framework to determine the targeted skills (i.e., position, reasons, and evidence) and levels and sequences of the activities within the scenario-based assessment.

Furthermore, we recognize that designing opportunities for students to demonstrate their content knowledge, skills, and abilities (KSAs) while minimizing construct-irrelevant variance is a complex process for EL students taking ELA tasks. Design procedures must incorporate attention to diverse student needs (Guzman-Orth, Laitusis, Thurlow, & Christensen, 2016; Pitoniak et al., 2009) as well as a deep understanding of how students learn and express their KSAs (Ketterlin-Geller, 2017) and how tasks can elicit relevant evidence so that the interpretations of the data can support the assessment claims (Mislevy, Steinberg, & Almond, 2003). To support this process, a multidisciplinary design team should also be a central feature of any task design, especially when designing tasks for ELs who are a diverse population with a variety of needs, especially linguistic needs (Solano-Flores, Shade, & Chrzanowski, 2014).

Two research questions guided this study: (a) *How do middle school ELs interact with an ELA task designed to measure argumentation skills in English?* and (b) *How can we improve the task design so that it increases the validity of the information that is elicited about EL students' argumentation KSAs?* Our research occurred in a multiphase process. We conducted cognitive labs with middle school EL students to understand their challenges in an argumentation assessment called Seaball—Semester at Sea. In the context of a fictional study abroad program, students need to demonstrate argumentation skills across five increasingly difficult activities (aligned to argumentation LPs). We also conducted a literature review to investigate how ELA content and assessment practices are made accessible for ELs. Finally, we synthesized information from the above two phases of work and made revisions to the Seaball task to improve access for middle school ELs.

Cognitive Lab Phase

Methods

Participants

Participants in the cognitive labs were from a convenience sample of eight seventh grade Spanish–English bilingual ELs (three females and five males) from an urban middle school in the mid-Atlantic region. Their overall ELP score determined from their district-administered ELP test ranged from beginning ($n = 3$), early intermediate ($n = 1$), and intermediate ($n = 4$; see Table 1 for student demographics).

Instruments

The instruments of the cognitive lab study included a background information questionnaire, the Seaball task, and a posttask survey.

Table 1 Student Demographic Characteristics and Seaball Performance Score

		English language proficiency test score levels						Content test score levels	
ID	Gender	Listening	Speaking	Reading	Writing	Comprehension	Overall	ELA	Math
ArgCL1	M	A	EI	I	I	I	I	Partial	Partial
ArgCL2	M	EA	I	EI	I	I	I	Not yet met	Partial
ArgCL3	F	EA	EA	EI	I	EI	I	Partial	Not yet met
ArgCL4	F	EA	I	EI	I	I	I		Partial
ArgCL5	M	EI	B	B	EI	B	B		Not yet met
ArgCL6	M	EI	B	EI	EI	EI	B		Partial
ArgCL7	M	EI	B	EI	EI	EI	B	Not yet met	Not yet met
ArgCL8	F	B	B	EI	EI	B	EI		Not yet met

Note. ID = unique identifier for each participant; M = male; F = female; A = advanced; EA = early advanced; I = intermediate; EI = early intermediate; B = beginner; ELA = English language arts test scores.

Background Information Questionnaire

The students' teacher filled out the background information questionnaire (BIQ), including student demographics such as gender, EL status, and standardized test scores for ELP, ELA, and math.

Seaball Task

In the Seaball Junk Food scenario, students engage in argumentation about whether junk food should be sold to students by interacting with game characters, evaluating claims and evidence, and making policy recommendations. Seaball Junk Food includes five activities of varying difficulty. In the first activity, Interview, students interview characters and classify their opinions into *ban* or *allow* categories. Next, students evaluate four candidates' profiles and then select an expert to address the students about junk food in the Select a Speaker activity. In the third activity, Identify Arguments, students identify the main claim, reasons, and evidence in the speech. Students then make a recommendation to the student council (i.e., ban or allow junk food) with supporting reasons in the Make a Recommendation activity. Finally, in the Establish a Criterion activity, students sort food items into junk food and healthy food categories, which involves using relevant evidence and evaluating people's arguments. The task is automatically scored, with 100 points possible.

Post-Task Survey

A 17-item online survey (12 Likert-type questions, five open-ended questions) elicited information about students' experience and perceptions of the Seaball task.

Procedure

Students participated in one-on-one cognitive labs with a trained researcher. Each cognitive lab lasted 60 – 90 minutes and was audio recorded. The students individually played through the Seaball task on the computer while researchers observed students' interactions (i.e., documenting issues with usability, language, and engagement). Researchers followed these observations with interview questions to learn more about students' perceptions and experience completing the task (e.g., difficult or unknown vocabulary, navigation issues). After completing Seaball, students filled out the post-task survey.

Results

Observation notes, interview transcripts, survey responses, and task log files (i.e., student actions and scores) were analyzed. Overall, these data sources indicated that although the students reported enjoying the experience of Seaball, the task was quite challenging for ELs. Notably, task scores were overall very low (range 27 – 78) due to students being unable to complete each activity before researchers helped them proceed to the next section due to time constraints. Specifically, five major themes emerged from qualitative analysis, related to issues with difficulty, usability, engagement, language, and timing.

Two primary sources of difficulty were related to task usability and linguistic complexity, including complexity of the directions as well as the content. Additionally, cultural accessibility posed an issue, with EL students being relatively unfamiliar with the specific context of studying abroad on a ship. Students were also unfamiliar with colloquialisms (i.e., junk food) and idiomatic expressions (i.e., jokes and humorous dialogue, a gamelike design element). The researchers had to help the students as they progressed through each activity in both usability and linguistic aspects (as a result, we recommend interpreting the student performance scores with some caution). These sources of difficulty also greatly affected students' time on task; students generally took extended time to decode and comprehend task directions, or they would try to "figure it out" through trial and error. Students were fatigued by the time they reached the end of the task. In general, EL students experienced difficulty in all aspects of Seaball. Specific issues for each Seaball activity are described below.

Activity 1: Interview

All students were engaged in this activity, but performance varied widely (range 0 – 100%; $M = 8/16$, or 50% correct; see Table 2). Most students needed assistance with the language, including help understanding the directions and the T-chart

Table 2 Summary of Observations for Activity 1: Interview

		Issues observed				Activity 1:
ID	Overall ELP	Difficulty	Usability	Engagement	Language	Classify an opinion (out of 15)[a]
ArgCL1	I	2	0	1	0	0 (0%)
ArgCL2	I	1	2	2	1	9 (60%)
ArgCL3	I	2	0	2	1	15 (100%)
ArgCL4	I	1	1	2	0	6 (40%)
ArgCL5	B	1	1	2	0	6 (40%)
ArgCL6	B	1	1	2	0	9 (60%)
ArgCL7	B	2	0	2	1	6 (40%)
ArgCL8	EI	0	2	2	2	12 (80%)
Mean score	–	1.25	0.88	1.88	0.63	8 (53%)

Note. ID = unique identifier for each participant; ELP = English language proficiency; 0 = no; 1 = partial; 2 = yes; I = intermediate; B = beginner; EI = early intermediate.

[a]Students did receive support from the interviewer in responding to the questions in the task.

graphic organizer (i.e., to classify opinions as ban/allow). Specific vocabulary was also troublesome, including construct-relevant (e.g., *opinion*, *ban, allow*), context-relevant (e.g., *junk food, concerned*), and general academic vocabulary (*at least*). Several students also showed some usability issues (i.e., using drag and drop). One student needed support during the entire activity.

Activity 2: Select a Speaker

Similarly, students were engaged in the second activity, but their performance indicated that they had difficulty with the task (range 35 – 75%; $M = 9.89$, or 50%; see Table 3). Almost all students needed help with usability; most reported confusion about how to proceed. This confusion could also be related to students' overall language difficulties, in addition to specific vocabulary (i.e., *junk, previous voyage, focus on this side of the debate, issue, buying*, and *include*). Three students also needed assistance navigating among the speakers' profiles (presented in tabbed format), and one had difficulty submitting a response. One student in particular (ArgCL7) had some difficulty with the typing required in the activity and was unable to produce an interpretable response.

Activity 3: Identify Arguments

Students were engaged and performed slightly better on this activity compared to the previous two activities (range 13 – 100%; $M = 10.38$, or 69%; see Table 4). Some students needed assistance with directions and did not understand what

Table 3 Summary of Observations for Activity 2: Select a Speaker

		Issues observed				Activity 2:
ID	Overall ELP	Difficulty	Usability	Engagement	Language	Select a speaker (out of 20)[a]
ArgCL1	I	1	1	1	1	7 (35%)
ArgCL2	I	0	1	2	0	10 (50%)
ArgCL3	I	2	0	2	1	9 (45%)
ArgCL4	I	0	2	2	1	15 (75%)
ArgCL5	B	1	1	1	0	9 (45%)
ArgCL6	B	1	1	2	0	7 (35%)
ArgCL7	B	0	2	1	2	9 (45%)
ArgCL8	EI	0	2	2	0	13 (65%)
Mean score	–	0.63	1.25	1.63	0.63	9.89 (50%)

Note. ID = unique identifier for each participant; ELP = English language proficiency; 0 = no; 1 = partial; 2 = yes; I = intermediate; B = beginner; EI = early intermediate.

[a]Students did receive support from the interviewer in responding to the questions in the task.

Table 4 Summary of Observations for Activity 3: Identify Arguments

		Issues observed				Activity 3:
ID	Overall ELP	Difficulty	Usability	Engagement	Language	Identify arguments (out of 15)[a]
ArgCL1	I	2	0	1	0	2 (13%)
ArgCL2	I	1	1	2	0	12 (80%)
ArgCL3	I	2	0	2	0	15 (100%)
ArgCL4	I	1	1	2	0	15 (100%)
ArgCL5	B	2	1	1	0	12 (80%)
ArgCL6	B	2	0	2	0	5 (33%)
ArgCL7	B	1	0	2	0	10 (66%)
ArgCL8	EI	2	0	2	0	12 (80%)
Mean score	–	1.63	0.38	1.75	0	10.38 (69%)

Note. ID = unique identifier for each participant; ELP = English language proficiency; 0 = no; 1 = partial; 2 = yes; I = intermediate; B = beginner; EI = early intermediate.

[a]Students did receive support from the interviewer in responding to the questions in the task.

to do. Some students did not know where to click on the screen to select the speaker's reasons. One student commented that there was a lot of information to synthesize and process, and another had spelling problems during typing. Two students in particular had difficulty with the English language in their typed responses. One student (ArgCL6) typed: "the soda and candys is bag for the salud of student" (*salud* is the Spanish equivalent of "health"). Another student (ArgCL7) showed reliance on phonetic spelling: "The yunk food is bad to de people and hte yunk food can be bad on the featuer."

Activity 4: Make a Recommendation

Overall, students performed in the mid-range again (range 33 – 100%; $M = 10.75$, or 60%; see Table 5). At this point, researchers reported that students were experiencing some fatigue and needed assistance to move forward.

Activity 5: Establish a Criterion

Overall, the data still indicated a wide range in performance (range 0 – 88%; $M = 16.50$, or 52%; see Table 6). All aspects of this section were difficult; interviewers reported that most students needed assistance throughout the activity to rephrase the directions, overcome usability issues (clicking in the correct areas), or rephrase the arguments characters provided.

Overall, results of the cognitive lab study identified multiple challenges faced by EL students taking a scenario-based argumentation skills assessment. To inform strategies to improve task accessibility, we next conducted a literature review.

Table 5 Summary of Observations for Activity 4: Make Recommendation

		Issues Observed				Activity 4:
ID	Overall ELP	Difficulty	Usability	Engagement	Language	Make a recommendation (out of 18)[a]
ArgCL1	I	2	0	1	0	14 (78%)
ArgCL2	I	2	0	2	0	14 (78%)
ArgCL3	I	2	0	1	1	6 (33%)
ArgCL4	I	0	2	0	0	18 (100%)
ArgCL5	B	1	1	2	0	6 (33%)
ArgCL6	B	2	0	1	0	12 (67%)
ArgCL7	B	0	0	0	0	8 (44%)
ArgCL8	EI	2	0	2	0	8 (44%)
Mean score	–	1.38	0.38	1.13	1.13	10.75 (60%)

Note. ID = unique identifier for each participant; ELP = English language proficiency; 0 = no; 1 = partial; 2 = yes; I = intermediate; B = beginner; EI = early intermediate.

[a]Students did receive support from the interviewer in responding to the questions in the task.

Table 6 Summary of Observations for Activity 5: Establish Criterion

		Issues observed				Activity 5:
ID	Overall ELP	Difficulty	Usability	Engagement	Language	Establish a criterion (score out of 32)[a]
ArgCL1	I	–	–	–	–	4 (13%)
ArgCL2	I	2	0	2	0	0 (0%)
ArgCL3	I	2	0	1	0	28 (88%)
ArgCL4	I	0	2	0	2	24 (75%
ArgCL5	B	2	0	2	0	12 (38%)
ArgCL6	B	2	0	2	0	18 (56%)
ArgCL7	B	2	1	1	0	18 (56%)
ArgCL8	EI	2	0	2	0	28 (88%)
Mean score	–	1.50	0.38	1.25	0.25	16.50 (52%)

Note. ID = unique identifier for each participant; ELP = English language proficiency; 0 = no; 1 = partial; 2 = yes; I = intermediate; B = beginner; EI = early intermediate.

[a]Students did receive support from the interviewer in responding to the questions in the task.

Literature Review Phase: Using Empirical Evidence to Inform Accessible Task Design

A literature review regarding ELA instruction, assessment, and accommodations for ELs was conducted to corroborate the information obtained from the cognitive labs. Combinations of search terms included "English learner, English language learner, English language learning, EL, and ELL" and "English language arts, ELA, language arts, argumentation" and "accessible, accessibility, accommodation, support." Academic journals as well as teacher handbooks and practitioner-focused websites were identified through major search engines (e.g., Google, Google Scholar, EBSCOhost).

Sources reviewed have indicated there is no one-size-fits-all approach to accessibility for ELs. Many means of input and output (e.g., oral and print language in English and students' home language, visuals, hands-on experiences) are recommended to allow ELs opportunity to access content and to demonstrate their KSAs, and these elements should be combined with clearly articulated learning goals, preteaching activities, multiple opportunities for practice, immediate feedback, collaborative learning with peers, frequent reteaching, and scaffolding (e.g., Gersten & Baker, 2000; Goldenberg, 2013; Robertson, 2017).

The EL accommodation literature has indicated that EL accommodations are typically designed to minimize linguistic construct-irrelevant variance that may impact the students' performance (although evidence points to the mixed efficacy of these supports; see Kieffer, Lesaux, Rivera, & Francis, 2009; Pennock-Roman & Rivera, 2011). Supports range from glossaries, including both in English and in students' home language (Cohen, Tracey, & Cohen, 2017; Graves, August, & Mancilla-Martinez, 2012; Solano-Flores et al., 2014; Wolf, Kao, Rivera, & Chang, 2012; Wolf, Kim, & Kao, 2012); read aloud supports (Buzick & Stone, 2014; Higgins & Katz, 2013; Higgins, Russell, & Hoffman, 2005; Laitusis, 2010; Wolf, Kao, et al., 2012), linguistic modification or use of plain English (Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005; Abedi & Lord, 2001; Abedi, Lord, Hoffsetter, & Baker, 2000; Abedi & Sato, 2007), and home language translations (Solano-Flores, 2006; Solano-Flores et al., 2014; Solano-Flores, Trumbull, & Nelson-Barber, 2002).

Best practices for assessment design have also offered guidelines that can improve student access. Elements such as Universal Design for assessment (Thompson, Johnstone, & Thurlow, 2002) or for learning (Center for Applied Special Technology, 2018) have emphasized the need to ensure that the language and any related visuals in the assessment are designed to minimize construct-irrelevant variance. Information should be clear, simple, and intuitive and presented in multiple modalities if it does not compromise measurement of the construct. Best practices for ELs have emphasized the importance of attending to student needs throughout the test development process, from conceptualization (e.g., clearly articulating the target population and use cases; Guzman-Orth, Lopez, Tolentino, Sova, & Stolow, 2016) to score reporting (Pitoniak et al., 2009). Placing the compilation of these practices within cognitive models for learning (Ketterlin-Geller, 2017) and evidence-centered design (Hansen & Mislevy, 2005, 2008; Mislevy et al., 2003) ensures that each step of the development process is conceptually and theoretically tied to the end goal to minimize the need for retrofitting the test for special populations, including expanded testing populations (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; International Test Commission,

2018; Wendler & Powers, 2009). Taken together, the evidence from our literature review suggests that these approaches are critical building blocks to implement at the initial stages of development for test design process.

Revision Phase: Using Empirical Evidence to Inform Accessible Task Design

In the final phase, we synthesized the findings from the data collection and the literature review. This synthesis indicated that the Seaball task could be heavily revised to better promote accessibility for ELs. Consistent with existing recommendations (e.g., Solano-Flores et al., 2014), we assembled a multidisciplinary review team consisting of experts in the construct, accessibility for ELs, and assessment design for ELs. The multidisciplinary review team designed and implemented revisions to the Seaball task, allowing for multiple rounds of review and consensus building from the team. Several revisions were considered for Seaball, but ultimately, the following changes were applied: reducing length, modifying the language, and adding supports (i.e., a glossary, read aloud instructions, a vocabulary-building activity, and visual cues). These revisions are detailed in the following sections (see Figure 1 for examples).

Activity Length

Most students had difficulty completing the Seaball assessment. To address this issue, four of the five activities were removed to focus on the first activity, Interview. The Interview activity content (i.e., identifying people's position on a controversial issue) is considered to be foundational in the argumentation LPs. The vocabulary introduced in this section is also pivotal for the subsequent activities. Thus, we now treat the first activity as a discrete activity that would allow students and teachers to start and stop this activity as needed, allowing for reteaching before proceeding to more complex activities in the Seaball assessment. With this change, it is important to recognize that only a portion of the argumentation construct is measured in the first activity compared to what can be measured with the entire Seaball assessment.

Linguistic Modification

We modified the language in the Seaball task according to guidelines that were developed and adapted from our synthesis of the literature and best practices for ELs (see Figure 1). Assessment specialists first consulted word lists for both the

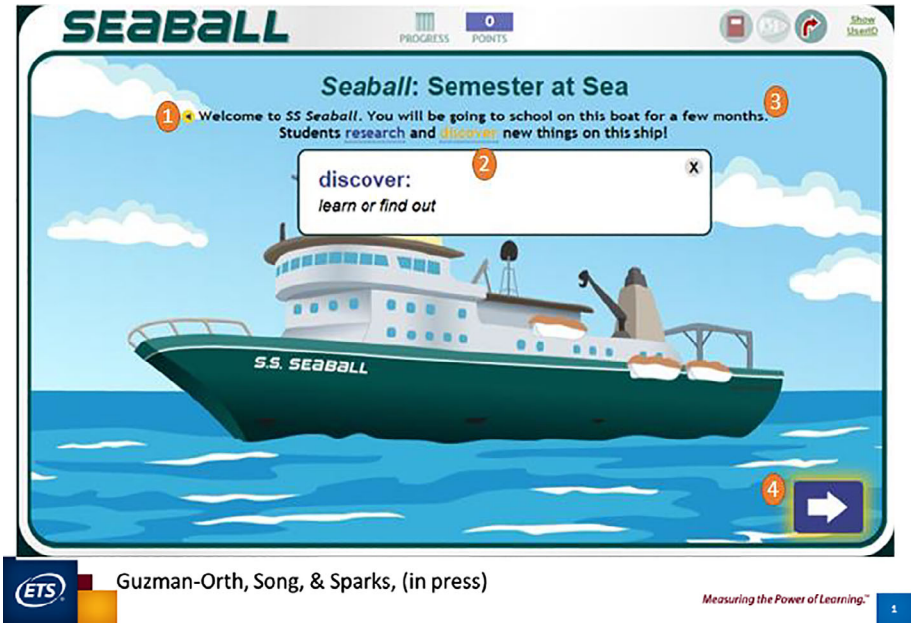

Figure 1 Revised introduction screen for the Seaball task. The redesigned supports for English language students are as follows: 1 = read aloud component for directions; 2 = pop up glossary (activated); 3 = linguistically modified text; 4 = enhanced visual cue to support task navigation. Not shown are the shortened task and vocabulary-building activity.

construct-relevant and context-relevant vocabulary to gain familiarity with the language that should not be modified (see appendix). Modification guidelines included the following: (a) modify construct-irrelevant language; (b) do not modify construct-relevant language; (c) do not modify the context-specific language (e.g., maintain the setting of the task in a fictional semester-at-sea program rather than adapting to a context that may be more culturally accessible); (d) use familiar/frequently used vocabulary for middle school ELs; (e) reduce sentence length; (f) use active voice and minimize passive voice; (g) shorten nominals/noun phrases; (h) reduce complex question phrases; (i) reduce comparative structures; (j) reduce prepositional phrases; (k) simplify sentence and discourse structure; (l) reduce subordinate clauses; (m) reduce conditional clauses; (n) reduce relative clauses; (o) use specific language rather than abstractions; (p) reduce negation; and (q) increase cultural accessibility (to support ELs who may not have prior knowledge of the semester abroad context). Taken together, these guidelines were applied to simplify the linguistic complexity of the task to an intermediate level of proficiency without changing measurement of the argumentation construct.

Glossary

An English glossary was added to assist in understanding terminology that was considered to be context relevant (i.e., to the semester-at-sea setting). The embedded, pop-up glossary in Seaball was designed to function similarly to the pop-up glossary supports in multistate consortia assessments (e.g., PARCC, Smarter Balanced), because students may have more familiarity with and exposure to this type of design. Similar to the linguistic modification guidelines, assessment specialists first consulted the appendix to gain familiarity with what context-relevant vocabulary could be glossed (because it was not removed in the linguistic modification step). Our approach to glossary guidelines comes from our synthesis and adaptation of existing guidelines in the field, such as (a) gloss any remaining words/phrases that are construct irrelevant; (b) gloss difficult vocabulary, including polysemous words (words with multiple meanings), false cognates, culturally relevant terms, idiomatic expressions, regional variations, etc.; (c) gloss words the first time they appear (i.e., subsequent words and plural forms do not need a gloss); (d) limit the number of words/phrases that are glossed on each page/screen (i.e., if there are multiple glossed words, consider modifying the language further to reduce the need for multiple glosses); and (e) gloss only relatively short words or phrases using concise, clear language and active voice.

Read Aloud

A read aloud (i.e., voice-over audio) component was added to the task directions. This modification provides multimodal input for students, especially those who may be struggling readers.

Vocabulary-Building Activity

We incorporated a vocabulary-building activity to ensure that students would have an opportunity learn key argument vocabulary prior to interacting with the task (i.e., *opinion, ban, allow, reasons*, and *evidence*). This activity was designed as a short quiz with immediate feedback to determine the students' prior knowledge of the key vocabulary and to correct misunderstandings. The quiz includes five three-option multiple choice questions. After each response, students receive immediate feedback and the correct definition. At the end of the quiz, students can review all five definitions.

Visual Cues

Because usability and navigation posed difficulties for the students, visual cues were added to help attract the students' attention to places on screen where they need to look or click next, including a glowing, pulsating yellow highlight appearing around buttons (see Figure 1). In sum, these principled revisions may mitigate the difficulties that ELs experienced.

Discussion

In this study, we investigated how ELs interacted with a scenario-based task measuring argumentation skills and how that task can be improved in a principled fashion to promote greater access for ELs to elicit evidence about their KSAs. Overall, data sources from a cognitive lab study (including observations, surveys, interviews, and performance data) overwhelmingly indicated that students enjoyed interacting with the task, but also experienced difficulties; these primarily involved

usability and linguistic considerations, which affected ELs' ability to independently access the content and demonstrate their knowledge and skills. A literature review to investigate best practices for ELA instruction and assessment for ELs, including accommodations, yielded important information to consider as well. Findings suggested that promoting access is not a one-size-fits-all approach, and that there is no single instructional activity or accommodation that works for all students, suggesting that a combination of approaches should be used. Critically, the revisions undertaken are not intended to make the content easier. Instead, the revisions are intended to minimize the construct-irrelevant language load.

Taken together, empirical findings from the cognitive labs and the literature review informed our approaches in the task revision phase. Our principled approach to designing EL supports for the Seaball task included removing all but one activity, adding the vocabulary-building activity to increase familiarity, adding visual cues to aid navigation, modifying the language and adding a glossary to support comprehension, and a read aloud component to help alleviate reading load and provide an additional means of access to task directions.

Study Limitations

There are some limitations to this work, specifically regarding the need for empirical evidence to support validation of the newly proposed design guidelines for accessible ELA tasks and their application to existing assessments. Currently, these guidelines have been applied to the Seaball tasks, but additional applications should be conducted with other task designs and other competencies and practices within the broader ELA construct (cf. Deane et al., 2015) to determine whether these guidelines are feasible for multiple purposes and to what extent the guidelines require adaptations to be more generally applied to other ELA tasks and skills. Additionally, evidence from students is needed to determine the effects of the principled revisions on students' performance to evaluate whether the revised version of the Seaball activity effectively minimizes the anticipated accessibility challenges for ELs. Such evidence is a necessary component for any validity argument supporting the use of this framework to inform future assessment task designs.

Directions for Future Research

Directions for future research include conducting a pilot test of the revised task to gather validity evidence of the principled approach to designing EL supports for tasks along the lines described above. Pending these results, similar revisions could also be applied to the remaining four Seaball activities.

Additional research should also include investigation of how classroom teachers implement the revised task as part of their formative assessment process. This implementation would provide necessary validity evidence to determine how teachers use Seaball, and Seaball's efficacy, as a tool for teachers to use when teaching argumentation skills to their students. It could also shed light on how students might use the information derived from Seaball. We anticipate that both ELs and non-ELs may benefit from the revised version of the first activity, and information on student performance from both groups would help to validate the revision decisions for the activity.

Further investigation of how ELs understand argumentation and demonstrate their reasoning is also needed. For example, ELs are learning argumentation content in English while they are also learning the English language. Some students in the pilot study used their home language to help demonstrate their knowledge, but students may not receive credit from teachers if responses are not in English or are misspelled. Using a different approach, such as a multilingual framework like *translanguaging* (the process of bilingual speakers using all of their linguistic resources in English and their home language to demonstrate meaning; see Canagarajah, 2006) and *conceptual scoring* (the process of giving credit to correct responses, regardless of what language is used; see Barrueco, Lopez, Ong, & Lozano, 2012), is recommended for bilingual speakers taking assessments measuring their functional KSAs, rather than their limited English proficiency (Guzman-Orth, Laitusis, et al., 2016; Guzman-Orth, Lopez, & Tolentino, 2017; Lopez, Guzman-Orth, & Turkan, 2015; Lopez, Turkan, & Guzman-Orth, 2017). In our study, students' responses to Seaball indicated that they partially understood the activity and could produce some relevant rationales (e.g., "the soda and candys is bag for the salud of student"). Despite typos and grammatical errors, the student stated that junk food (soda and candy) is bad for students' health (*salud* is the Spanish equivalent of "health"). The phonetic spelling in the response, "The yunk food is bad to de people and hte yunk food can be bad on the featuer," indicated that, typos aside, the student had internally contextualized the content of the task and was demonstrating knowledge that junk food can be harmful. This student's understanding included the use of language at the *morphemic level*, the smallest, meaningful units of language (e.g., phonemes): The students' use of the

"y" in *yunk [*sic*]* mirrored their accent and pronunciation of the /j/ English phoneme, which sounds like the "*ll*" letter in Spanish that is pronounced like "y" ("*j/yah*"). In these cases, although students were having some difficulties throughout the various components of the activities, they were actively seeking ways to meaningfully participate and demonstrate their knowledge using all of their linguistic resources.

This study provides an example of how multiple sources of data can be synthesized to provide evidence to inform assessment design decisions. Despite widespread attention to evidence-based practices for EL accessibility on large scale assessments, less attention has been focused on EL accessibility for classroom-based assessments or activities that can be used as part of the formative assessment process to support teachers and students in monitoring learning. Additionally, this study is unique in that it focuses on accommodating EL students' language needs on ELA tasks—an effort that is highly complex due to the integrated language demands of the content, the need to maintain the measurement of the argumentation construct, students' ongoing acquisition of the English language, and the need to provide language supports for their ongoing language acquisition. We hope that this study can increase attention to the diverse needs for EL students to promote accessibility on ELA tasks so that those ELA tasks can yield more valid, actionable results for teachers and students to monitor and support learning.

References

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification* (CSE Report 666). Retrieved from http://cresst.org/publications/cresst-publication-3037/

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14,* 219–234. https://doi.org/10.1207/S15324818AME1403_2

Abedi, J., Lord, C., Hoffsetter, C., Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26. https://doi.org/10.1111/j.1745-3992.2000.tb00034.x

Abedi, J., & Sato, E. (2007). *Linguistic modification—A report prepared for the US Department of education LEP partnership*. Washington, DC: US Department of Education.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Barrueco, S., Lopez, M., Ong, C., & Lozano, P. (2012). *Assessing Spanish-English bilingual preschoolers: A guide to best approaches and measures*. Baltimore, MD: Brookes.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, *8*(2–3), 70–91.

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

Buzick, H. M., & Stone, E. A. (2014). A meta-analysis of research on the read aloud accommodation. *Educational Measurement: Issues and Practice, 33*(3), 17–30. https://doi.org/10.1111/emip.12040

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly, 3*(3), 229–242.

Center for Applied Special Technology. (2018). UDL guidelines version 2.2. Retrieved from http://www.udlcenter.org/aboutudl/udlguidelines

Cohen, D., Tracey, R., & Cohen, J. (2017). On the effectiveness of pop-up English language glossary accommodations for EL students in large-scale assessments. *Applied Measurement in Education, 30*(4), 259–272. https://doi.org/10.1080/08957347.2017.1353986

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*. Retrieved from http://www.corestandards.org/the-standards/ELA-Literacy

Deane, P., Sabatini, J., Feng, G., Sparks, J. R., Song, Y., Fowles, M., . . . Foley, C. (2015). *Key practices in the English language arts: Linking learning theory, assessment, and instruction* (Research Report No. RR-15-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12063

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

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*, 577–589. https://doi.org/10.1037/a0014702

Gersten, R., & Baker, S. (2000). What we know about effective instruction practices for English-language learners. *Exceptional Children, 66*, 454–470. https://doi.org/10.1177/001440290006600402

Goldenberg, C. (2013). Unlocking the research on English learners. What we know – and don't yet know – about effective instruction. *American Educator*, *37*(2), 4 – 11.

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools – A report to the Carnegie Corporation of New York*. Washington, DC: Alliance for Educational Progress.

Graves, M. F., August, D., & Mancilla-Martinez, J. (2012). *Teaching vocabulary to English language learners*. New York City, NY: Teachers College Press.

Guzman-Orth, D., Laitusis, C., Thurlow, M., & Christensen, L. (2016). *Conceptualizing accessibility for English language proficiency assessments* (Research Report No. RR-16-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12093

Guzman-Orth, D., Lopez, A. A., & Tolentino, F. (2017). *A framework for the dual language assessment of young dual language learners in the United States*. (Research Report No. RR-17-37). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12165

Guzman-Orth, D., Lopez, A. A., Tolentino, F., Sova, L., & Stolow, A. (2016, December). *Considerations in assessing dual language learners*. Paper presented at the California Education Research Association, Sacramento, CA.

Hansen, E. G., & Mislevy, R. J. (2005). Accessibility of computer-based testing for individuals with disabilities and English language learners within a validity framework. In M. Hricko & S. Howell (Eds.), *Online assessment and measurement: Foundation, challenges, and issues* (pp. 214 – 262). Hershey, PA: Idea Group. https://doi.org/10.4018/978-1-59140-720-1.ch011

Hansen, E. G., & Mislevy, R. J. (2008). *Design patterns for improving accessibility for test takers with disabilities* (Research Report No. RR-08-49). Princeton, NJ: Educational Testing Service.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive process in writing* (pp. 3 – 30). Hillsdale, NJ: Lawrence Erlbaum.

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* Washington, DC: Council of Chief State School Officers.

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, *28*(3), 24 – 31. https://doi.org/10.1111/j.1745-3992.2009.00151.x

Higgins, J., & Katz, M. (2013). A comparison of audio representations of mathematics content. *Journal of Special Education Technology, 28*(3), 59 – 66. https://doi.org/10.1177/016264341302800305

Higgins, J., Russell, M., & Hoffman, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *The Journal of Technology, Learning, and Assessment, 3*(4), 1 – 35.

Hillocks, G., Jr. (2002). *The testing trap: How state writing assessments control learning*. New York, NY: Teachers College Press.

International Test Commission. (2018). ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations. [www.InTestCom.org]

Ketterlin-Geller, L. M. (2017). Understanding and improving accessibility for special populations. In A. A. Rupp and J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and application*. Hoboken, NJ: John Wiley & Sons.

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *79*, 1168 – 1201. https://doi.org/10.3102/0034654309332490

Kuhn, D. (1991). *The skills of argument*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511571350

Laitusis, C. C. (2010). Examining the impact of audio presentation on tests of reading comprehension. *Applied Measurement in Education, 23*, 153 – 167. https://doi.org/10.1080/08957341003673815

Lopez, A. A., Guzman-Orth, D. A., & Turkan, S. (2015). How might a translanguaging approach in assessment make tests more valid and fair for emergent bilinguals? In G. Valdés, K. Menken, & M. Castro (Eds.), *The common Core and English language learners/emergent bilinguals: A guide for all educators* (pp. 266 – 267). Philadelphia, PA: Caslon. https://doi.org/10.1002/ets2.12140

Lopez, A. A., Turkan, S., & Guzman-Orth, D. (2017). *Conceptualizing the use of translanguaging in initial content assessments for newly arrived emergent bilingual students* (Research Report No. RR-17-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12140

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3 – 62. https://doi.org/10.1207/S15366359MEA0101_02

National Center for Education Statistics. (2012). *The nation's report card: Writing 2011* (NCES 2012-470). Washington, DC: Institute for Education Sciences, U.S. Department of Education.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, *30*(3), 10 – 28. https://doi.org/10.1111/j.1745-3992.2011.00207.x

Perkins, D. N., Farady, M., & Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 83 – 106). Hillsdale, NJ: Erlbaum.

Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., & Ginsburgh, M. (2009). *Guidelines for the assessment of English language learners*. Princeton, NJ: Educational Testing Service.

Robertson, K. (2017). *Supporting ELLs in the mainstream classroom: Language tips. Color in Colorado WETA public broadcasting*. Retrieved from http://www.colorincolorado.org/article/supporting-ells-mainstream-classroom-language-tips

Shore, J. R., Wolf, M. K., & Blood, I. (2013). *ELFA teacher's guide*. Retrieved from https://www.ets.org/s/research/pdf/elfa_teachers_guide.pdf

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teachers College Record, 108*(11), 2354 – 2379. https://doi.org/10.1111/j.1467-9620.2006.00785.x

Solano-Flores, G., Shade, C., & Chrzanowski, A. (2014). *Smarter Balanced Assessment Consortium: Item accessibility and language variation conceptual framework*. Retrieved from https://portal.smarterbalanced.org/library/en/item-accessibility-and-language-variation-conceptual-framework.pdf

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, *2*, 107 – 129. https://doi.org/10.1207/S15327574IJT0202_2

Song, Y., Deane, P., & Fowles, M. E. (2017). *Examining students' ability to critique arguments and exploring assessment and instructional implications* (Research Report No. RR-17-16). Princeton, NJ: Educational Testing Service.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (NCEO Synthesis Report No. 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Wendler, C., & Powers, D. (2009). What does it mean to repurpose a test? *R&D Connections, 9*. Princeton, NJ: Educational Testing Service.

Wolf, M. K., Kao, J. C., Rivera, N. M., & Chang, S. M. (2012). Accommodation practices for English language learners in states' mathematics assessments. *Teachers College Record*, *114*(3), 1 – 6.

Wolf, M. K., Kim, J., & Kao, J. (2012). The effects of glossary and read-aloud accommodations on English language learners' performance on a mathematics assessment. *Applied Measurement in Education*, *25*, 347 – 374. https://doi.org/10.1080/08957347.2012.714693

Appendix

EL Argumentation: Construct-Relevant Vocabulary List

Construct Relevant Vocabulary (in order of appearance in the task)

Argumentation/argument/argue
Evidence
Reason/reasoning
Opinion/point of view/viewpoint
Agree/disagree
Concern/concerned
In my opinion/I think/I believe
Controversy

Context Relevant Vocabulary (in order of appearance in the task)

Ship	Cooler
Semester at Sea	Notebook
Voyage	Map
Bridge	Library
Student council	Cafeteria
President	Student lounge
C-Store	Fitness center
Junk food	Hallway
Interview	Snacks
Ban/allow	Captain
Meeting room	Characters
Calorie	

Suggested citation:

Guzman-Orth, D., Song, Y., & Sparks, J. R. (2019). *Designing accessible formative assessment tasks to measure argumentation skills for English learners* (Research Report No. RR-19-15). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12251

Action Editor: Don Powers

Reviewers: Paul Deane and Maria Elena Oliveri