frontiers in Education

ORIGINAL RESEARCH
published: 04 September 2019
doi: 10.3389/feduc.2019.00088

Considerations for Designing Accessible Educational Scenario-Based Assessments for Multiple Populations: A Focus on Linguistic Complexity

*Maria Elena Oliveri**

Educational Testing Service, Princeton, NJ, United States

As the diversity of the test-taker population increases so should assessment development practices evolve to consider the various needs of the multiple populations taking the assessments. One need is the ability to understand the language used in test items and tasks so they do not present unnecessary challenges for the test-takers, which may be misconstrued as a lack of knowledge of the content assessed. This investigation is important because linguistic complexity may constitute a source of construct-irrelevant variance, which may render an assessments' passages and questions less accessible and present unnecessary challenges to the multiple test-taker populations potentially leading to score misinterpretation, disengagement with the task, and increased cognitive load. To develop more linguistically accessible assessments for multiple populations, less accessible construct-irrelevant text may require modification and less accessible construct-relevant text may need scaffolding. In this paper, I discuss considerations for designing accessible assessments for multiple populations with a focus on linguistic complexity. To illustrate these considerations, I refer to digitally delivered scenario-based tasks of English Language Arts framed in a science context.

Keywords: English learner, formative assessment, culturally and linguistic diversity, scenario - based learning, text complexity

OPEN ACCESS

Edited by:
Mustafa Asil,
University of Otago, New Zealand

Reviewed by:
Guillermo Solano-Flores,
Stanford University, United States
Raman Grover,
British Columbia Ministry of Education, Canada
Jacqueline Evers-Vermeul,
Utrecht University, Netherlands

****Correspondence:***
Maria Elena Oliveri
moliveri@ets.org

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *02 March 2019*
Accepted: *02 August 2019*
Published: *04 September 2019*

Citation:
Oliveri ME (2019) Considerations for Designing Accessible Educational Scenario-Based Assessments for Multiple Populations: A Focus on Linguistic Complexity. Front. Educ. 4:88. doi: 10.3389/feduc.2019.00088

INTRODUCTION

The diversity of test-taker populations is increasing given globalization, immigration, and the rising cultural and linguistic diversity of the examinees. An example are English Learners (ELs), who are one of the fastest growing populations in the United States. In the last decade, as compared to 7% growth of the general student population, the EL population increased by 60%. Demographers project that by 2025; one out every four students attending public schools will be classified as an EL (Hodgkinson, 2008). Concomitant with changes in student demographics are advances in technology and the increased use of technology-enhanced items in assessments. These shifts and innovations present opportunities for new assessment designs and technologies. One such opportunity is to advance the conceptualization and design of technology-enhanced educational assessments administered to multiple populations. For instance, given the diversity of the multiple populations' familiarity with the item formats, language used in the test, or the technology used in the assessment, developers may need to expand the traditional approaches used to investigate fairness.

Test fairness review processes typically involve analyzing items by investigating differential item functioning (DIF) posttest administration. Ercikan and Oliveri (2013) and Oliveri et al. (2014) identified shortcomings related to DIF analyses when tests are administered to heterogeneous populations, including under detecting DIF, suggesting the need to expand review processes from analyzing DIF post-administration to considering fairness starting from the conceptualization and design of assessments.

To augment traditional test fairness approaches, a sociocognitive approach to test development has been proposed to consider early on (during assessment design) the needs of multiple test-taker populations (International Test Commission, 2018; Mislevy, 2018). The sociocognitive approach calls developers to attend to key elements of task design, construct representation, and the type of resources and knowledge culturally and linguistically diverse populations might bring to the assessment (Weir, 2005; O'Sullivan and Weir, 2011; Turkan and Lopez, 2017; Mislevy, 2018). One of the central goals of the approach is to guide decisions (e.g., the types of language, vocabulary, and visual representations) that can be included in an assessment without creating unnecessary construct-irrelevant variance (CIV) to allow for valid and fair score-based interpretations when assessing populations from diverse backgrounds. According to the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), CIV is "variance in test-taker scores that is attributable to extraneous factors that distort the meaning of the scores and thereby decrease the validity of the proposed interpretation" (p. 217).

Complementing a sociocognitive assessment development approach are conceptual frameworks such as the Universal Design for Learning (UDL), which describe ways to develop assessments that are more accessible to diverse populations. Kettler et al. (2018) define accessibility as the extent to which a product (i.e., a test) eliminates barriers and permits equal use of components or services for diverse populations. UDL is defined as "an approach that involves developing assessments for the widest range of students from the beginning while maintaining the validity of results from the assessment" (Thurlow et al., 2010, p. 10). "Universal design" describes a movement within architecture that aims to design buildings to accommodate the widest range of users, including individuals with disabilities (Rose and Meyer, 2000; Rose and Strangman, 2007). Architects who apply the principles of universal design consider the multiple needs of potential users during the design stage, to avoid the expensive and often awkward retrofitting of buildings after construction (Dolan et al., 2005). In 1997, the Center for Universal Design (CUD) formally defined universal design as "the design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design" (CUD, 1997, p. 1).

In an application of UDL to assessments and learning activities, Rose and Meyer (2000) describe that UDL principles allow flexible access to test accommodations in online tasks to meet students' needs more effectively. Multiple representations of items can be included such as multiple media (text, video, audio), highlighters, and zoom magnification. Russell et al. (2009) indicate that a computer-based system that employs the principles of UDL is one that: "[r]equires developers to build features into the architecture of a system that allow accommodation tools to be accessed flexibly to meet the needs of each individual user. In a universally designed test-delivery system, all students across a testing program use the same standard interface and have access to high-quality tools and accommodations delivered in a controlled, standardized, and equitable manner" (p. 3).

Boals et al. (2018) suggest that although UDL principles were designed to apply to the widest range of learners, their focus on cognitive neuroscience approaches to learning makes them more useful to students with disabilities than culturally and linguistically diverse populations. The authors suggest complementing UDL's cognitive neuroscience focus with a sociocultural framing. The goal is to make UDL principles be more inclusive of diverse populations' language-related needs.

Building on a sociocognitive approach to assessment design (Mislevy, 2018), UDL and recent guidelines (e.g., International Test Commission, 2018), this paper addresses two main objectives. First, it identifies language-related considerations relevant to the design of more linguistically accessible assessments for multiple populations. This objective is important because the use of unnecessarily complex language may reduce the accessibility of assessments administered to diverse populations and reduce test fairness and opportunity to learn. Second, because language affects test takers' ability to demonstrate what they know and can do, approaches to addressing text complexity, avoiding construct-irrelevant linguistic complexity, and scaffolding construct-relevant linguistic complexity when assessments are administered to multiple populations are described. To address these objectives, the Green Islands, an example of an English Language Arts (ELA) scenario-based assessment (SBA) is used as a case study. This discussion is important now—at a time of increasing demographic diversity both nationally and internationally.

CONSIDERATIONS RELATED TO TEXT COMPLEXITY

Text complexity can be defined at various levels of the text (e.g., vocabulary, syntax, explicitness of relations between clauses/sentences/paragraphs, information density, level of abstractness). When thinking about text complexity, it is important to differentiate language that is central to the assessed content (construct-relevant language) from language that is not (construct-irrelevant language). Examples of construct-relevant language include the linguistic demands related to the content of the assessment such as the use scientific terms and content-related rhetorical structures. An example of construct-irrelevant language would be the use of overly complex language in general test directions.

As Avenia-Tapper and Llosa (2015) explain, the classification of what language to modify and what to maintain at a higher level of complexity is a departure from previous work on assessments of content knowledge. Earlier approaches assumed complex linguistic features are a source of CIV by virtue of their complexity. Instead, Avenia-Tapper and Llosa propose matching an assessments' linguistic features with the domain (or construct) targeted by the assessment to inform linguistic modifications without under representing the construct. In science, these considerations involve providing students with opportunities to learn science-related complex linguistic structures represented in scientific texts and talk.

Considerations of text complexity are important because depending on the assessment goals and the construct assessed, text complexity may or not be desirable and/or constitute a source of CIV. In reading comprehension tests, CIV is introduced by using language that is above the level of proficiency intended by the test due for instance to: (a) vocabulary that is above grade-level, (b) using too many technical terms, or (c) a failure to allow sufficient time to read passages (for tests intended to be unspeeded). In content-related tests, such as mathematics or science, a heavy reliance on reading comprehension skills may be a source of CIV. The inclusion of CIV in ELA assessments may threaten the validity of score-based inferences (Haladyna and Downing, 2004; Oliveri and von Davier, 2016) and reduce students' opportunity to learn.

The existence of CIV in ELA assessments is illustrated by research that shows that the size of the gap between populations such as (English Learners) ELs' and non-ELs' performance on standardized tests can be reduced when the language used in test questions is modified. Abedi et al. (2003) found that ELs performed between 40 and 60% lower than non-ELs in ELA assessments. However, the performance gap was substantially smaller (8–25%) after modifying (simplifying) the items' linguistic demands. Abedi (2006) posits that: "By reducing the impact of language barriers on content-based assessments, the validity and reliability of assessments can be improved, resulting in fairer assessments for all students" (p. 381). Thus, he suggests modifying unnecessarily construct-irrelevant complex language used in test questions. He suggests basing such modifications on the knowledge of content/linguistic experts and the actual characteristics of test items.

Acknowledging the importance of reducing CIV in assessments administered to multiple populations and addressing text complexity when designing assessments, Guzman-Orth et al. (2016) explain that designing assessments that are fair, meaningful, and accessible for diverse populations is a multistep and complex process. An important step is to use language that is at the right complexity level for the intended test-taker population. Another is to embed scaffolds to support student learning. These steps require careful analysis to identify which text to modify (Abedi, 2008; Sato, 2008) and which one to scaffold and how. However, these decisions require tradeoffs and considerations of various factors such as the short- and long-term effects of linguistic complexity and linguistic modification.

For instance, consideration of linguistic complexity needs to involve various factors including test fairness as well as the tradeoff between the short- and long-term goals of improving literacy. In the short term, there might be a discrepancy in reading ability between students with higher/lower reading proficiency. While it is possible to decrease text-level demands to render materials easier to read for lower-performing readers, some of which may be ELs to increase test scores, in the longer term, this might create additional difficulties in relation to the actual preparation of students to handle real-world texts that are complex and not modified.

Further, as T. O'Reilly (personal communication, February 21, 2019) points out, the Common Core (which is an educational initiative that details what K—12 students throughout the United States should know in ELA and mathematics at the conclusion of each school grade) advocates for proficiency in grade and out of grade level texts. Therefore, reducing text-level complexity does not help with the "out of grade level" aim of the Common Core. Moreover, the Common Core advocates for content area and disciplinary literacy, both of which are likely to increase text-level demands by including technical vocabulary. Thus, the overarching problem becomes one of improving short-term reading outcomes while remaining attentive to long-term growth goals related to properly preparing students who will have to read increasingly more complex texts in the real world. One plausible answer that requires additional empirical research is how to include text-level supports and increasingly more complex text over time, and how to gradually reduce the level of supports and scaffolding presented to students through appropriate fading mechanisms. Beyond reducing the supports provided through the online learning tools, one would also have to reduce the supports provided in the classroom to better prepare students over time so that students can handle more complex text when there is no support outside the classroom.

As an example, in the domain of reading, one skill students need to learn is how to make inferences. While in the short-term, making the text more explicit is beneficial to help students understand the text, students also need to be provided with situations that afford them with opportunities to learn the skill of drawing inferences. To this end, students thus need to interact with increasingly more challenging texts to learn how to draw inferences across various sets of grade-level texts. Consequently, appropriate supports that are gradually faded away can be included in the online learning tasks to help students become increasingly less dependent on modified and scaffolded text. It is important to keep these goals in mind when designing and developing digitally delivered assessments (T. O'Reilly, personal communication, February 21, 2019).

RESEARCH ON LINGUISTIC COMPLEXITY AND ITS CONSEQUENCES

Previous research has been conducted to investigate linguistic complexity as a possible source of CIV in assessments using various approaches including textual analysis and think-aloud protocols among others (Martiniello, 2008; Sato et al., 2010; Young et al., 2014). The studies have been conducted to investigate possible ways to modify test items without modifying the assessed construct and to identify possible consequences associated with the presence of CIV due to text complexity on cognition and test-taking behavior (Abedi, 2006; Martiniello, 2008; Turmo and Elstad, 2009).

Martiniello (2008) conducted textual analysis of math word problems and think-aloud protocols using data from the Massachusetts Comprehensive Assessment System fourth-grade math test. Her study revealed construct-irrelevant language in items related to the use of multiple clauses, complex structures, long noun phrases, limited syntactic transparency (i.e., lack of clear relationships between the syntactic units), and unfamiliar vocabulary. Examples of vocabulary-related linguistic complexity included items that had more than one unknown word in each sentence, the inclusion of words that were long, had multiple meanings, or were morphologically complex.

Young et al. (2014) used data from K-12 content assessments to investigate linguistic modification of 120 test items in mathematics and science administered to 4th and 6th grade students that were identified as having a wide range of outcomes for ELs and non-ELs on the item performance between the original and modified versions of the items. No systematic differences were found in relation to item performance for either group. However, the study did point to 11 categories relevant to linguistic modification such as:

- Removing empty context or information that makes the items context less direct and making the context more explicit;
- Simplifying challenging words that are not content related and replacing them with more accessible words;
- Unpacking the complex ideas provided in the items;
- Reducing wordiness in the overall item and ensuring that the item's stem is clear;
- Reducing the use of *if* clauses and breaking sentences down into simpler sentences;
- Changing the use of the passive voice to the active voice and using the present tense of verbs more frequently than past, future, or conditional tenses;
- Reducing the use of extraneous words;
- Emphasizing key words by underlining them; and
- Reducing the use of unnecessary visuals, graphics, or artwork.

The study conducted by Young et al. (2014) was an expansion of the work conducted by Sato et al. (2010), which only contained five modification categories (context, graphics, vocabulary/wording, sentence structure, and format/style).

Examples of Consequences of Using Complex Construct-Irrelevant Language

Abedi (2006) and Martiniello (2008) suggest that the use of complex construct-irrelevant language in items may disadvantage some populations. For instance, students may not understand what the items ask or may require extra time to read and comprehend them. In such cases, students may respond incorrectly. Moreover, the use of such language may prevent students from building mental representations of what they read (Sheehan et al., 2014). Students may also disengage because the language of the task is too complex and the task is not within their zone of proximal development (Vygotsky, 1978).

The use of unduly complex language also may increase second language leaners' cognitive load with detrimental effects to their learning, as suggested by findings from a study conducted by Turmo and Elstad (2009). The study used data that examined the linguistic factors in items (e.g., item wording, vocabulary familiarity) contained in the Grade 5 and 8 standardized Oslo 2007 science tests. Results revealed that culturally and linguistically diverse populations performed lower than native speakers for test questions that contained unfamiliar vocabulary and technical terms. The authors explain that one reason is that overly complex language places too many demands on culturally and linguistically diverse populations' cognitive load (Paas et al., 2003). The reasons may be that students need to attend to too many cognitive activities including decoding, comprehending, and understanding the science content; all of which they have yet to master.

The use of unnecessarily complex language may also result in the misinterpretation of students' scores. For instance, teachers may incorrectly attribute low performance to a lack of content mastery rather than lack of understanding of unfamiliar vocabulary, technical terms, or unduly complex language. Misinterpretations may occur because culturally and linguistically diverse students may have had less exposure to regular instruction or have a home language that differs from the school language (International Test Commission, 2018). In other instances, a low score may indicate decreased student engagement with the task (Wiliam, 2011; Snyder, 2016). In these various cases, teachers may misjudge students' ability to handle competently more advanced content knowledge. They may also fail to provide increasingly more challenging content to students on the subject of the test, which may lead to missed opportunities to learn and further disadvantage culturally and linguistically diverse populations (Abedi and Lord, 2001; Abedi and Gándara, 2006; Kieffer et al., 2009).

Approaches to Reduce the Use of Complex Construct-Irrelevant Language in Assessments

Given the abovementioned consequences associated with using complex construct-irrelevant language, which sometimes disadvantages some populations, researchers have investigated the kinds of linguistic features that might help render item language more transparent and accessible for diverse populations (Cummins et al., 1988; Abedi et al., 1997, 2000; Abedi, 2014). Alderson (2000) and Coltheart (1981) suggest using concrete words to assist readers build mental images to facilitate reading comprehension. Abedi et al. (1997) suggest rewording mathematics word problems to make semantic relationships more explicit without affecting the underlying content structure. Thus, the reader is more likely to correctly construct a representation of a word problem and solve it satisfactorily. The authors also suggest reducing wordiness and the number of clauses used in a sentence, adding emphasis to key words, or representing words graphically. They also suggest replacing unfamiliar vocabulary with more frequent synonyms or simplifying verb forms.

Additional suggestions on how to write linguistically accessible questions for diverse populations come from guidelines. Examples include the Guidelines for the Assessment of English Language Learners (Pitoniak et al., 2009); the ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations (International Test Commission, 2018); and Abedi (2008) and Sato (2008) Guide to Linguistic Modification: Increasing English Language Learner Access to Academic Content. These documents guide test developers on ways to address the challenges that may occur when designing tests that are fair to enable valid-score based inferences to be made when assessing culturally and linguistically diverse populations.

Among the suggestions provided by the guidelines are to write items that contain clear language that is at the appropriate level for the intended test taker populations. Such language includes the use of simple sentence structures and familiar topics and contexts. It also includes the avoidance of regional or sensitive vocabulary, words with multiple meanings, language that is ambiguous, or complex words or passages that are not construct-relevant. The guidelines also suggest using test instructions that maximize clarity, and selecting common topics, scenarios, figures, and images that are neither offensive nor derogatory to any cultural or language group or that may cause an emotional reaction. The considerations outlined in test development guidelines for diverse populations and associated research on linguistic features were used when developing the Green Islands SBA, which I used as a case study, as elaborated later in the article.

CONSTRUCT-RELEVANT LINGUISTIC COMPLEXITY

Definition and Examples

As T. O'Reilly (personal communication, February 21, 2019) points out, however, not all complex text needs to be modified or simplified because such decisions may also reduce diverse learners' opportunity to learn and acquire new strategies to learn how to understand construct-relevant complex text. Construct-relevant text complexity relates to linguistic demands that align to the assessed construct. In a science task, the inclusion of new terms that may be unfamiliar or sentence structures that emulate science talk are needed to provide students with opportunities to learn construct-relevant language. Thus, Turkan and Lopez (2017) suggest that the unknown terms, which relate to the (science) content of the test, do not need modification.

In relation to maintaining the needed level of linguistic complexity, Burstein et al. (2012) discuss the need for students to gain exposure to grade-level curriculum, which often involves interacting with reading materials that are complex. If students read easier texts or work with easier materials, they do not interact with text at the complexity level the educational standards require. To this end, scaffolding of construct-relevant language may be needed to support student learning through technology-enhanced assessments.

Scaffolding to Support the Acquisition of Construct-Relevant Complex Language

The use of instructional scaffolds can facilitate students' acquisition of construct-relevant language. The scaffolds may also help students access grade-level content to reach targeted learning goals while developing their English proficiency (Echevarria et al., 2004). The use of scaffolds may also help learners be better prepared to understand less accessible text and its features to be better equipped to read similar texts in the future (Burstein et al., 2012; Wolf et al., 2016). For example, if teachers instruct students about how to understand prefixes and suffixes, they are better preparing students to understand how to make sense of other prefixes and suffixes encountered in other texts. I provide additional examples of such scaffolds later in the article in reference to the Green Islands SBA.

Wolf et al. (2016) provide examples of scaffolds to support instruction when using technology-enhanced items. These include using immediate feedback, story retelling, and optional prompting. Boals et al. (2018) also suggest the use of other scaffolds such as illustrating the types of dialogic interactions peers have at school when discussing content-related issues with each other or with teachers and experts.

Language-Related Considerations When Developing Scenario-Based Assessments

O'Reilly and Sabatini (2013) describe that SBAs are a cluster of techniques for organizing and sequencing a set of thematically-related reading passages, sources, and items in a digital (online, computer-delivered) environment. As compared to traditional print reading, SBAs support the integration of multiple types of reading materials such as reading passages, websites, or documents on a single device. This design enables SBAs to be more useful assessment approaches than traditional assessments as SBAs more closely align to today's literacy demands where readers have access to multiple reading materials within a single device.

Accordingly, SBAs include items to assess students' ability to read strategically, to achieve specific goals, and to evaluate the importance and relevance of the materials read. Students are provided with an overarching goal for reading thematically related sources to solve problems, make decisions, apply strategies, or complete a more complex task. Tasks that are more complex may include writing a summary, a letter to students' parents, or making a presentation. Students may also need to respond to questions about the materials read. Questions may include traditional comprehension items (e.g., identify key information), or less traditional ones, which may involve asking students to synthesize or integrate multiple texts, evaluate web search results, complete graphic organizers, or apply what they read to a new situation or context. In SBAs, tasks are sequenced to build up students' knowledge and help them develop a deeper model of the content. Reading strategies such as the use of graphic organizers help students build mental models. These strategies are designed to model good practices and support learning.

Simulated "peer" students are included in SBAs to guide students, provide feedback, or model ways to solve a problem.

The SBA may include support features and scaffolds to provide opportunities to students to learn new content, clarify information, or acquire new vocabulary. For instance, SBAs provide opportunities for students to learn to use language communicatively while performing academic tasks (Wolf et al., 2016).

Beyond the abovementioned advantages, despite their more complex and elaborate design, SBAs have several advantages as compared to traditional assessments (e.g., based on multiple-choice discrete items). SBAs provide richer details about the context of the problems or scenarios presented in the tasks that comprise the content of the SBAs (Bennett, 2016). SBAs also are designed to increase opportunities for students to engage in richer, deeper, and more meaningful opportunities to learn an expanded set of skills. SBAs may allow educators to better track student learning, understand where breakdowns occur in their learners' thought processes, and provide ways to support diverse students who underperform.

The flexibility of SBAs affords benefits to all students and may be particularly useful in helping to expand educational opportunities in underserved areas (such as low-income or rural communities, or when students have reduced access to tutors). The development of linguistically accessible SBAs, which is needed when the tasks are used with multiple populations in alignment with the earlier mentioned sociocognitive approach to assessment development may involve considering which text to modify/simplify and which to scaffold.

These goals are important because although SBAs are innovative and have various advantages as compared to traditional forms of assessment, research is needed to identify approaches to tailor them to the needs of diverse populations. This research is important given the demographic changes occurring in schools in the United States concomitant with advances in technology and the learning sciences.

CASE STUDY

Instrument Used to Exemplify Analysis of Text Complexity of Scenario-Based Assessments

To analyze the above mentioned considerations related to construct-relevant and construct-irrelevant text complexity and language-related scaffolding issues, I will describe the linguistic analyses conducted when developing the SBA referred to as the Green Islands, which is used as a case study. "The Green Islands" is a modular SBA designed to support the teaching and learning of reading literacy skills with a theme-based content related to a science context for third-grade students. The science content used in the SBA is related to the Next-Generation Science Standards (Achieve Inc, 2013) topic of biological evolution (unity and diversity). Accordingly, the items and reading passages ask students to read texts about various ways in which animals survive and adapt (e.g., through camouflage), as well as read about the characteristics of animals living in different habitats and their requisite needs. The choice to design the literacy task in a science-related context is consistent with the goals described in National Research Council (2014), which indicate the need to address students' difficulties related to reading and understanding science texts by providing students with opportunities to interact with various types of reading materials (e.g., newspapers or web content). Given that the goal of this section of the article is to use Green Islands to illustrate the text-related complexity considerations discussed earlier, the discussion that follows exemplifies the implications of text-complexity considerations on instrument development. To this end, in what follows, I provide examples of reading supports and scaffolds that can be included in SBAs to provide students with opportunities to learn technical vocabulary and content that might be difficult to acquire through other forms of assessment. First, I will describe the Green Islands. Then, I will illustrate the evaluation of the Green Islands' linguistic complexity.

The Green Islands SBA

The Green Islands' content aligns to the Common Core State Standards for English Language Arts (Common Core State Standards Initiative, 2013). Therefore, the learning objectives and items assess skills listed in the Common Core State Standards. Examples of skills that align to the Common Core State Standards include asking and answering questions to demonstrate their understanding of a text; referring explicitly to the text as the basis for the answers, or determining the main idea of a text; and recounting the key details and explaining how they support the main idea. Moreover, the Green Islands is based on the Cognitively Based Assessments of, for, and as Learning (CBAL) Building and Sharing Knowledge (B&SK) key practices (O'Reilly et al., 2015). The skills in the CBAL B&SK key practices include subskills that students can acquire gradually by engaging in thematically related texts. Examples include setting goals and activating prior knowledge, understanding the text, clarifying meanings, consolidating, and conveying knowledge. Therefore, the activities in the Green Islands ELA task have been carefully selected to provide opportunities to students to practice these skills and subskills. Thus, the SBA's flexible design serves as an environment to allow for organizing and sequencing thematically related types of information, content, and items.

At the beginning of the Green Islands task, students are informed that they won the Science Explorers Contest and, as their prize, they won a free trip to the Green Islands to conduct science. Once on the Green Islands, they will learn about the animals living on the islands, their habitats, the characteristics that enable them to adapt to their environment, and the weather on the islands. Students will meet scientists, ask questions, gather information, and write summaries to share with their parents and classmates back home that which they have learned. Information is presented in various communication modalities including simulated dialogs, chats, and text formats.

Students interact with the task in various ways, such as by writing information in their notebook. Students are provided with opportunities to learn vocabulary in context, use graphic organizers, and complete increasingly challenging items. Throughout the task set, the students interact with simulated peers and scientists who offer guidance and tips. The items are designed to provide learners with opportunities to learn foundational concepts that will be introduced in the Green Islands and to help them acquire unfamiliar content-related vocabulary and content. The selection of materials to include as easier items was informed by consultation with third grade teachers and discussion with subject matter experts who pointed out the types of words, language use, and content that may be challenging for third-grade students and would need modification or scaffolds depending on whether the terms were construct-relevant or -irrelevant. The inclusion of adaptive paths also assists learners to work at a level that better matches their zone of proximal development (Vygotsky, 1978). In such cases, students may be more engaged with the task (Snyder, 2016).

Text Complexity Evaluation

The language-related considerations discussed next are not meant to fit all SBA development efforts but provide more concrete examples and suggestions than those provided in guidelines (e.g., International Test Commission, 2018), which are typically developed to apply to more general test development efforts than the development of SBAs. Such suggestions may not be applicable to all SBAs, which may assess other constructs, for other populations, or for use in other contexts.

To inform linguistic modifications and scaffolding of the Green Islands SBA, the assessment developers consulted with subject matter experts and used natural language processing (NLP) tools. Subject matter experts (e.g., teachers and experts with a background in second language learning, school psychology, science, and linguistics) provided insights to the assessment development team regarding whether the language used in the task was grade-level appropriate, whether they thought students would face particular struggles with the technology, avatars, or scaffolds used in the task. They also provided feedback on the ways (formats and types of representations) in which the skills were captured, and the appropriateness/suitability of the samples of the literacy passages and activities included in the SBA. Moreover, they provided insights regarding the results of the NLP tools to evaluate the SBA's language.

NLP tools were also used to evaluate the complexity of the text in the SBA (e.g., by grade-level) with respect to various linguistic features such as cohesion, syntax, and vocabulary difficulty. NLP tools can help evaluate the reading passages and text by evaluating various linguistic features such as word concreteness, word unfamiliarity, academic vocabulary, syntactic complexity, lexical cohesion, argumentation, and narrativity against previously established metrics tested on large numbers of texts and reading passages (cf. Sheehan et al., 2015). I elaborate on each of these approaches later in the article.

These approaches were used to evaluate the following questions:

1. Is the text in the Green Islands at, above, or below grade-level?
2. What language features of the text are too simple or too complex for the targeted grade level?
3. Is the text flagged as too simple or complex construct-relevant or construct-irrelevant? Are the sentences or language well-defined and specific or vague and general? How well do the sentences stand alone to complete an idea?
4. What are useful approaches to modify the construct-irrelevant complex language?
5. What are possible ways to scaffold the construct-relevant language?

Evaluation of Text Complexity Using Natural Language Processing Tools

TextEvaluator^{TM} was used to evaluate the level of text complexity of the Green Islands' items and passages. TextEvaluator is a comprehensive text-analysis system designed to help teachers, textbook publishers, test developers, and literacy researchers select reading materials that are consistent with the text-complexity goals outlined in the Common Core State Standards (Sheehan et al., 2014, 2015). TextEvaluator utilizes NLP methodologies to provide unbiased estimates of the complexity level (i.e., grade level) of text passages and identify whether the text is below, within, or above the grade-level of the targeted standards.

TextEvaluator is an ETS-developed tool that analyzes linguistic features of texts such as vocabulary. This evaluation involves using a standardized frequency index (SFI) to calculate the difficulty of vocabulary. The SFIs are calculated based on the frequency of word occurrences in the Gigaword 4 corpus (Parker et al., 2009). It is one of the largest corpus collections for English comprised of 9.8 million newswire articles. Words that have a high SFI are regarded as very frequent and therefore very familiar, even to young students. In contrast, words with lower SFIs occur less frequently in the English language and are considered more difficult.

To evaluate text complexity, NLP results were combined with the Spache readability Formula (Spache, 1953). The Spache Formula is used for primary-grade texts to calculate the difficulty of the text, based on the word unfamiliarity and sentence lengths of the sentences in a text. Words that readers encounter frequently are likely to be familiar requiring less cognitive energy and time to interpret.

Expert Reviewers' Evaluation of Text Complexity

Beyond the above mentioned insights provided by expert reviewers, the experts also provided suggestions regarding the extent to which the instructions used in the SBA were clear. They also evaluated the results of the NLP tools such as vocabulary that had low SFIs and made suggestions related to whether to modify/scaffold particular language or terms. For instance, reviewers were asked to examine the vocabulary to evaluate the extent to which it was at, above, or below grade-level. To conduct this evaluation, reviewers were provided with grade-level appropriate and construct-relevant examples of key terms and complex vocabulary (e.g., endotherm, ectotherm, camouflage, habitats) not needing modification or simplification because they are topic- or content-relevant as exemplified in unit-related Next Generation Science Standards documents (Achieve Inc, 2013). On the other hand, reviewers were also presented with examples of complex or wordy instructions that can be simplified as findings from the abovementioned research noted that such language can render the items less clear and add unnecessary complexity to the task (Sato et al., 2010; Young et al., 2014).

During the text-complexity-evaluation process, reviewers also were asked to identify vocabulary or phrases that might make the item or passage hard to understand. The objective was to help uncover any aspects (e.g., the phrasing and context) of the items' and passages' language that students may have a hard time understanding. Along the same lines, reviewers were asked to think about vocabulary or phrasing that made items or passages hard or confusing to answer. These questions helped to identify the types of modifications that could be made to help clarify the items and passages for students of diverse linguistic backgrounds; see below for specific questions reviewers examined to evaluate text complexity.

Evaluation and Modification of Construct-Irrelevant Language

Construct-irrelevant terms were modified using the aforementioned guidelines. The checklist provided in **Table 1** provides a summary of the linguistic features that were evaluated as they may lead to CIV when assessments are administered to CLD populations. Examples are the use of vocabulary that may have multiple meanings or may be confusing.

Evaluation and Scaffolding of Construct-Relevant Language

Scaffolds were developed in the Green Islands to provide students with opportunities to learn construct-relevant terms. As described earlier, one of the advantages of SBAs is their flexible design space. The design enables the inclusion of scaffolds to provide learners with opportunities to practice task-related literacy skills.

TABLE 1 | Modification checklist for construct-irrelevant language guided by the ITC guidelines (2018).

Context

- Does the item contain topics or concepts that may be considered offensive, derogatory, or exclusionary to the targeted culture?
- Is the item free of references to historic events, product names, entertainers, geography, government, holidays, measurement systems, or currency that are not central to construct and unfamiliar to the targeted populations?

Vocabulary and language use

- Does the item contain the use of ambiguous language that may not be readily understood by diverse populations?
- Can the vocabulary lead to a different meaning in another language [or cultural group]?
- Is the idea contained in the item synonymous between languages [and cultures]?
- Are there multiple meanings for words contained in the item?
- Does the item contain words that are unnecessarily complex?
- Does the item contain regional and sensitive vocabulary that may be unfamiliar for diverse populations?

Instructions

- Are the test instructions simple and clear?

RESULTS

Results of NLP Tools Used to Evaluate Text Complexity

Figure 1 shows the results of the analysis of the different words from the Green Islands SBAs. The figure shows that most words had an SFI >60. Hence, most words were likely to be familiar to students in third grade.

Table 2 shows the ranked order of a subset of Green Islands' words and the results of the TextEvaluator analysis. Words that have an SFI <50 are more difficult for young readers. For instance, words such as *and*, *the*, *of*, and *to* are very common as compared to less frequently found words such as *endotherm*, *ectotherm*, *reptile*, *tortoise*, and *iguana*,

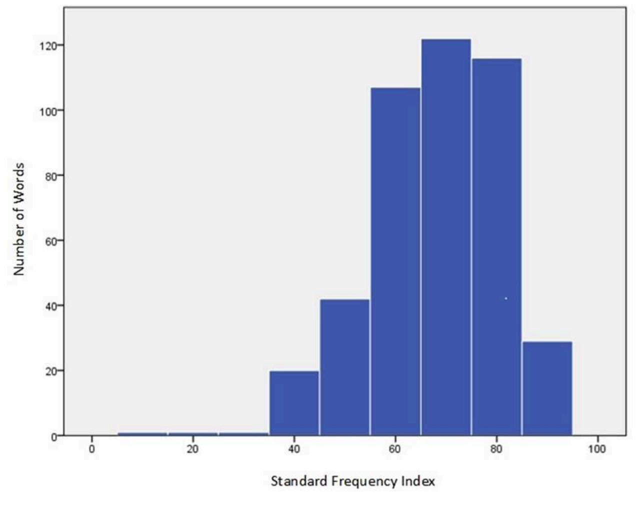


FIGURE 1 | Analysis of familiarity of words in the Green Islands' task.

TABLE 2 | Results of word frequency analysis.

SFI	No. of words	Sample words
0–10	3	Ectotherm, endotherm
11–20	4	Ectotherms, endotherms
31–40	30	Stubby, gills, anatomical, iguanas, tortoises, reptile, claws
41–50	92	Adaptation, offspring, explorers, crabs, turtles, penguins
51–60	279	Rocks, wings, birds
61–70	359	Living, ready, probably
71–80	429	Some, all, people
81–90	247	And, the, of, to
Total	1,443	

which are the more difficult words found in the Green Islands SBA.

Results of Expert Reviewers' Evaluation of Text Complexity

After the task's text was run through TextEvaluator, experts reviewed the items and resulting SFI results. The SFI results helped inform linguistic modifications and scaffolding decisions depending on whether the terms were construct-relevant or irrelevant. Examples of construct-relevant words included technical terms used in the task to convey science-related knowledge.

Modification of Construct-Irrelevant Language

Experts determined that words, such as *probably* and *congratulations* were construct-irrelevant. Such words were unfamiliar and were modified. They were replaced with more familiar, higher frequency words. An example of a word with a SFI below 50 is given in **Table 3**. The words "congratulations" is hard to sound out and can be modified with two shorter words "good news".

Sentence length and sentence structure

Results of the TextEvaluator analysis also pointed out sentences that might be long or unclear. As noted in the examples provided in **Table 4**, the developers identified ways to break down the sentences into one or more to increase readability, clarity, and cohesion without jeopardizing the content. As a reviewer pointed out, the sentence could have been made more cohesive by including connectives and/or cue phrases.

TABLE 3 | Example of words with SFI below 50 and proposed modifications.

Original: Congratulations. You are a winner of this year's Science Explorers contest.

Revision: Good news. You are a winner of this year's Science Explorers contest.

Note: Three features suggest that "congratulations" is a complex word: (a) it has five syllables, (b), it is not on the Spache list of familiar words, and (c) it has an SFI below 50.

TABLE 4 | Example of sentences with low cohesion and proposed modifications.

Original: I am a tortoise. Some people call me a turtle. But I am not a turtle. I live on land. I have a heavy, round shell. Turtles live in the water. Most turtles have light, flat shells.

Revision: I am a tortoise. Some people call me a turtle. But I am not a turtle. I live on land. Most turtles live in the water. I have a heavy round shell. Most turtles have light flat shells.

Note: The original sentence has low cohesion because there is minimal word overlap across sentences. To increase cohesion, the order of the two sentences was switched.

TABLE 5 | Example of revised instructions to enhance clarity for test takers.

Original: Next, you are going to learn some interesting facts about the needs and wants of animals. You will learn about…

- Animals' needs: What animals must have to survive.
- What animals do to survive.
- The kinds of places animals live in or their habitats.

Revision: Next, you are going to learn about animals' needs, characteristics, and habitats.

a. Animals' needs: What animals must have to survive.
b. Animals' characteristics: The special features that help animals survive.
c. Animals' habitats: How the places that animals live in help them survive.

Test instructions

The International Test Commission (2018) Guidelines suggests designing test instructions to maximize clarity (i.e., use simple and clear language). It also suggests that test developers or publishers should provide evidence that the language used in the test instructions and test items is clear for the test takers. After the experts' review of the Green Islands task, revisions were made to be the original set of instructions to clarify the language. In the example shown in **Table 5**, the instructions were revised to make the three bullet points parallel in structure. As a reviewer pointed out, further revisions could have included reducing the repetition of complex words to reduce the overall amount of text students read, turning the questions into real questions, and avoiding ambiguous words such as "them" in the final sentence.

Scaffolding Construct-Relevant Language

Experts found the names of the animals living in the Green Islands (e.g., crabs, tortoises, and penguins) and the types of adaptations (endotherm and ectotherm) to be construct-relevant. Moreover, the results of the TextEvaluator analysis flagged words, such as colonies, habitats, and research station as possibly unknown or unfamiliar to some test takers. Because these words are construct-relevant, they would not be modified. Instead, scaffolds were designed to support learners' acquisition of complex but construct-relevant terms. Consistent with UDL, a multimedia approach was used to scaffold students' learning of the unknown terms. Hence, words, images, and photos were used for the new terms.

Figure 2 shows an example of pop-up illustration glossaries, which are pictorial representations of words or terms displayed on the computer's screen for students to click on demand. In the Green Islands, students can view the habitats they will encounter (i.e., forest, volcanoes, meadow, or waterfall) by clicking on the red dot next to the term in question. The pairing of text and visuals may reinforce learning, as long as the cognitive load resulting from processing the information from two channels and then integrating it is not increased too much (Rose and Meyer, 2000). **Figure 3** illustrates the Green Islands' research station. **Figure 4** illustrates the term "habitats." An interactive activity was included to provide students with an opportunity to learn the names of the habitats (rainforests, meadows, arid zones, and beaches) they would learn about in the task. **Figures 2–4** provide information to students on the locations they will visit and help orient students to the task's sites they will view on the Green Islands.

Figures 5, **6** show scaffolds designed to provide opportunities to students to learn the morphology of words they encounter in the task (e.g., breaking words apart by roots, suffixes, and prefixes). This approach is applied to learning words about animal types such as endotherm and ectotherm and the adaptations that have occurred to the animals to accommodate temperature differences. Teaching learners about morphological structure can contribute to their understanding of the unknown words they may encounter in the future (Kieffer and Lesaux, 2007).

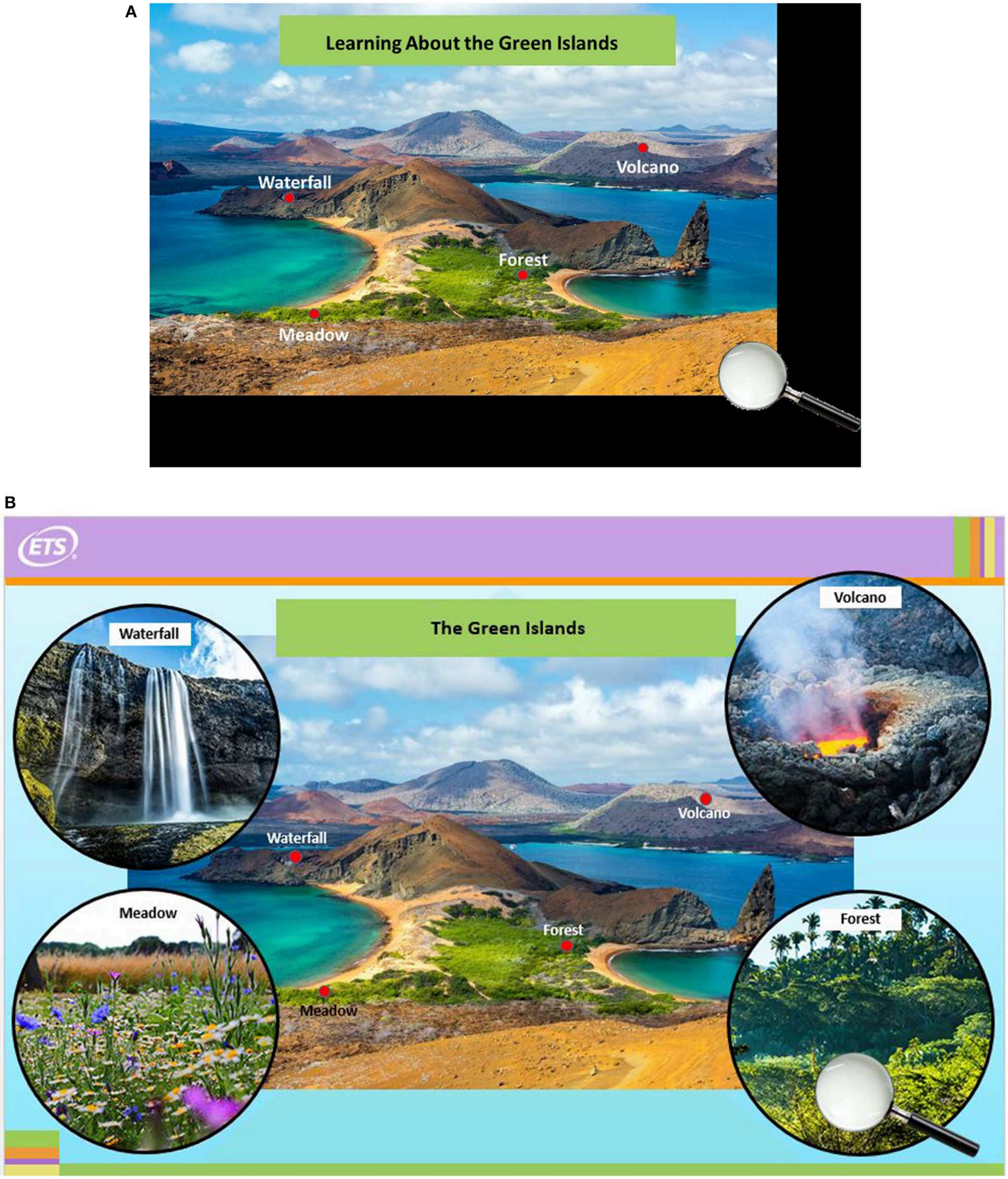


FIGURE 2 | (A) Example of a pop-up illustration glossary. ©Michaël Lejeune, CC-BY-SA-3.0, Wikimedia Commons. **(B)** Example of a pop-up illustration glossary. ©Michaël Lejeune, CC-BY-SA-3.0, Wikimedia Commons.

DISCUSSION

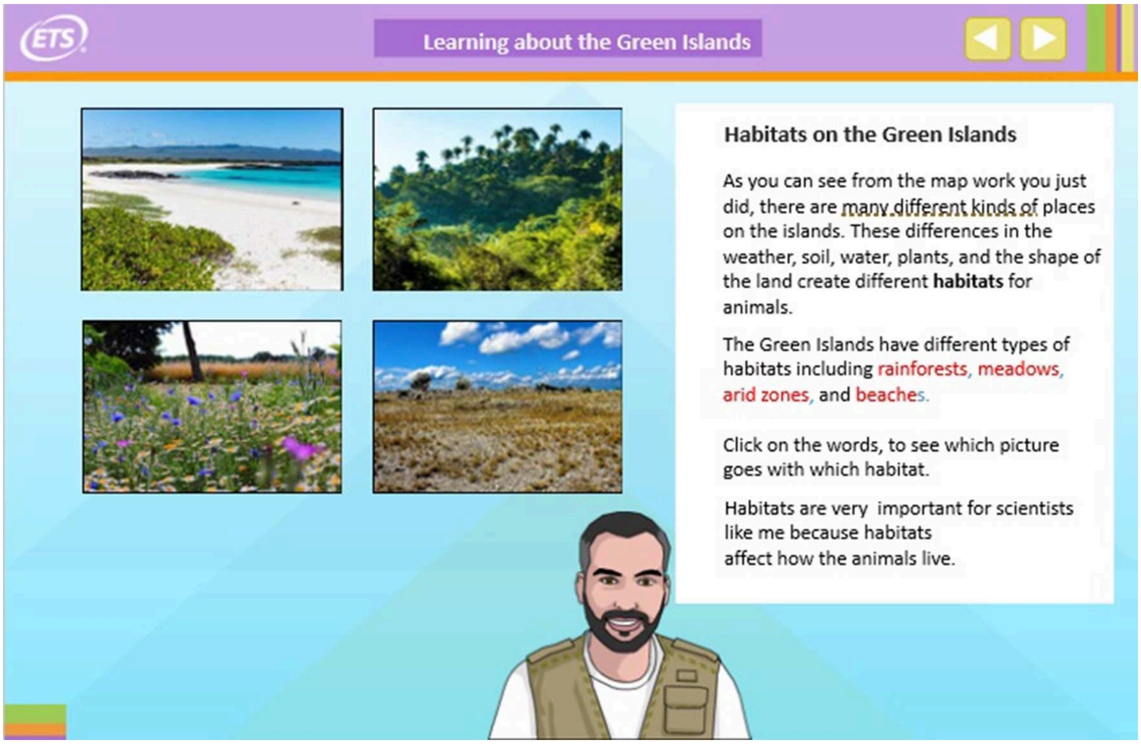
In this study, I discussed language-related considerations relevant to designing SBAs for multiple populations and suggested various approaches to evaluating text complexity, modifying construct-irrelevant language, and scaffolding construct-relevant language. I described steps that go beyond traditional test fairness review processes to consider the use of accessible language in the design phases of test development. Moreover, I illustrated the use of NLP tools and expert reviews to evaluate text complexity of text in ELA SBAs.

FIGURE 3 | Example of a visual of the Green Islands Research Station.

These suggestions were informed by previous research, guidelines, and UDL principles that collectively aim to enhance fairness, validity, and accessibility of items for diverse populations. These documents and tools were used to illustrate an approach to develop more accessible tasks for linguistically diverse populations.

I illustrated the approach using a third-grade, scenario-based ELA assessment contextualized in science. I argued for controlling text complexity at the design stage of SBAs to avoid unnecessary retrofitting post-SBA development. This focus on design is not new and has been discussed earlier in evidence-centered design (Mislevy et al., 2003), UDL (Rose and Meyer, 2000), and the more recent sociocultural focus that extends UDL principles for diverse populations (Boals et al., 2018). These authors highlight the importance of understanding the characteristics of the focal populations to be assessed. This understanding is needed to support meaningful score-based interpretations based on the use of more accessible ELA tasks.

Although the process presented may have advantages, such as identifying suggested steps for controlling text complexity, and supporting multiple populations' learning of construct-relevant terms; it also has limitations. One limitation is that the proposed approach has not yet been tested out with students. New challenges may emerge when the tasks and scaffolds are tried out with students in pilot studies. Such studies would need to be conducted to examine how well the scaffolds work with diverse populations and to determine the



Learning about the Green Islands

Habitats on the Green Islands

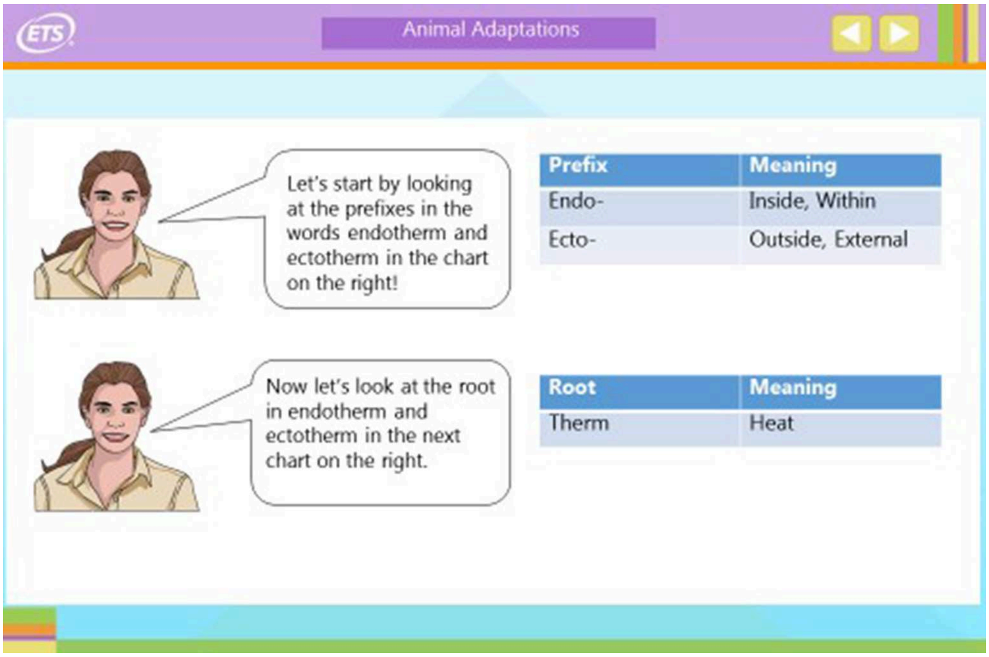
As you can see from the map work you just did, there are many different kinds of places on the islands. These differences in the weather, soil, water, plants, and the shape of the land create different **habitats** for animals.

The Green Islands have different types of habitats including rainforests, meadows, arid zones, and beaches.

Click on the words, to see which picture goes with which habitat.

Habitats are very important for scientists like me because habitats affect how the animals live.

FIGURE 4 | Example of an interactive activity to learn the habitats on the Green Islands.



Animal Adaptations

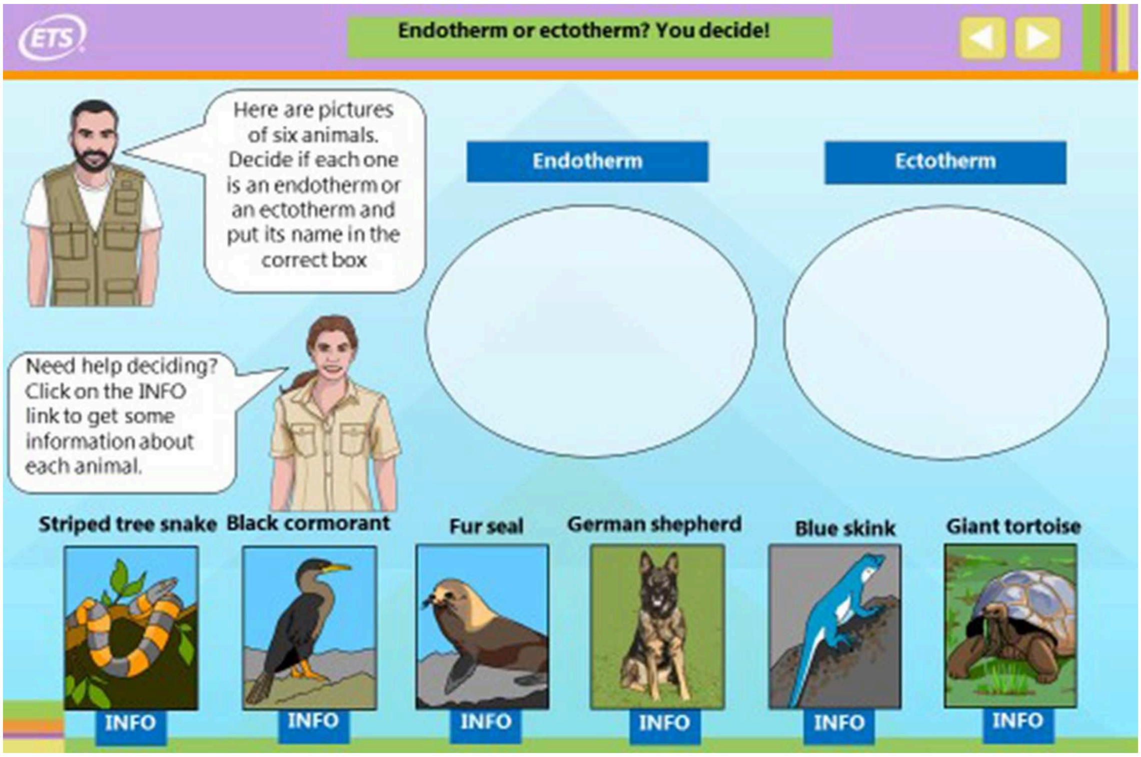
Let's start by looking at the prefixes in the words endotherm and ectotherm in the chart on the right!

Prefix	Meaning
Endo-	Inside, Within
Ecto-	Outside, External

Now let's look at the root in endotherm and ectotherm in the next chart on the right.

Root	Meaning
Therm	Heat

FIGURE 5 | Example 1 of scaffolding activities to learn construct-relevant vocabulary.



Endotherm or ectotherm? You decide!

Here are pictures of six animals. Decide if each one is an endotherm or an ectotherm and put its name in the correct box

Need help deciding? Click on the INFO link to get some information about each animal.

FIGURE 6 | Example 2 of scaffolding activities to learn construct-relevant vocabulary.

generalizability of findings to other contexts, content areas, and SBAs. Therefore, this study described an approach that is potentially helpful in revising items and questions in SBAs to increase linguistic accessibility of items for diverse populations; however, additional research is needed with the populations targeted for assessment.

The scaffolds included in the Green Islands task are designed to support linguistically diverse learners' acquisition of foundational literacy skills in a science task. Additional research is needed to identify the extent to which the selected scaffolds work. Specific questions to address include scaffolds' use (e.g., are students using the scaffolds?), timing (e.g., are scaffolds presented early enough?), and type (e.g., are there other scaffolds that might work better to help diverse learners better comprehend the content and context of the task?). Various approaches, such as NLP tools, process data on students' use of the scaffolds, cognitive laboratories, or a combination of these approaches may be used to conduct the investigations. This research is needed to inform the development of educational SBAs administered to multiple populations. These questions and further research is important at a time when science and the language used to express it is evolving and the populations taking ELA and science assessments are rapidly changing.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

The author thanks Kathleen Sheehan, Paul Borysewicz, Maria Konkel, Rene Lawless, and Theresa Richter for their assistance on the Green Islands project funded by ETS.

REFERENCES

Abedi, J. (2006). "Language issues in item-development," in *Handbook of Test Development,* eds S. M. Downing and T. M. Haladyna (Mahwah, NJ: Erlbaum), 377–398.

Abedi, J. (2008). "Part I: language factors in the assessment of English language learners: the theory and principles underlying the linguistic modification approach," in *Linguistic Modification* (Washington, DC: LEP Partnership), 2–52. Available online at: https://ncela.ed.gov/files/uploads/11/abedi_sato.pdf

Abedi, J. (2014). The use of computer technology in designing appropriate test accommodations for English language learners. *Appl. Meas. Educ.* 27, 261–272. doi: 10.1080/08957347.2014.944310

Abedi, J., and Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educ. Meas.* 25, 36–46. doi: 10.1111/j.1745-3992.2006.00077.x

Abedi, J., Leon, S., and Mirocha, J. (2003). *Impact of Students' Language Background on Content Based Assessment: Analyses of Extant Data* (CSE Tech. Rep. No. 603). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., and Lord, C. (2001). The language factor in mathematics tests. *Appl. Meas. Educ.* 14, 219–234. doi: 10.1207/S15324818AME1403_2

Abedi, J., Lord, C., Hofstetter, C., and Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educ. Meas.* 19, 16–26. doi: 10.1111/j.1745-3992.2000.tb00034.x

Abedi, J., Lord, C., and Plummer, J. (1997). *Language Background as a Variable in NAEP Mathematics Performance* (CSE Tech. Rep. No. 429). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Available online at: https://cresst.org/wp-content/uploads/TECH429.pdf

Achieve Inc (2013). *Third Grade. Next Generation Science Standards*. Available online at: https://www.nextgenscience.org/sites/default/files/3%20combined%20DCI%20standardsf.pdf (accessed August 7, 2019).

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Avenia-Tapper, B., and Llosa, L. (2015). Construct relevant or irrelevant? The role of linguistic complexity in the assessment of English language learners' science knowledge. *Educ. Assess.* 20:95. doi: 10.1080/10627197.2015.1028622

Bennett, R. E. (2016). *Opt Out: An Examination of Issues* (Research Report RR-16-13). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12101

Boals, T., Castro, M., and Willner, L. S. (2018). "Moving beyond assumptions of cultural neutrality to improve accessibility and opportunity to learn for English language learners," in *Handbook of Accessible Instruction and Testing Practices,* eds S. N. Elliot, R. J. Kettler, A. P. Beddow, and A. Kurz (Cham: Springer), 119–134.

Burstein, J., Shore, J., Sabatini, J., Moulder, B., Holtzman, S., and Pedersen, T. (2012). *The Language Muse System: Linguistically Focused Instructional Authoring* (Research Report No. RR-12-21). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2012.tb02303.x

Center for Universal Design (CUD) (1997). *About UD: Universal Design Principles*. Available online at: http://www.webcitation.org/5eZBa9RhJ (accessed August 7, 2019).

Coltheart, M. (1981). The MRC psycholinguistic database. *Quart. J. Exp. Psychol.* 33, 497–505.

Common Core State Standards Initiative (2013). *English Language Arts Standards*. Available online at: http://www.corestandards.org/ELA-Literacy/ (accessed August 7, 2019).

Cummins, D. D., Kintsch, W., Reusser, K., and Weimer, R. (1988). The role of understanding in solving word problems. *Cogn. Psychol.* 20, 405–438. doi: 10.1016/0010-0285(88)90011-4

Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., and Strangman, N. (2005). Applying principles of universal design to test delivery: the effect of computer-based read aloud on test performance of high school students with learning disabilities. *J. Technol. Learn. Assess.* 3:7. Available online at: https://ejournals.bc.edu/index.php/jtla/article/view/1660 (accessed August 7, 2019).

Echevarria, J., Vogt, M., and Short, D. (2004). *Making Content Comprehensible for English Learners: The SIOP Model* 2nd ed. Boston, MA: Pearson/Allyn and Bacon.

Ercikan, K., and Oliveri, M. E. (2013). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations," in *Validity and test use: An International Dialogue on Educational Assessment, Accountability and Equity,* ed M. Chatterji (Bingley: Emerald Publishing), 69–86.

Guzman-Orth, D., Laitusis, C., Thurlow, M., and Christensen, L. (2016). *Conceptualizing Accessibility for English Language Proficiency Assessments* (Research Report No. RR-16-07). Princeton, NJ: Educational Testing Service.

Haladyna, T. M., and Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educ. Meas.* 23, 17–27. doi: 10.1111/j.1745-3992.2004.tb00149.x

Hodgkinson, H. (2008). *Demographic Trends and the Federal Role in Education*. Center on Education Policy. Available online at: files.eric.ed.gov/fulltext/ED503865.pdf

International Test Commission (2018). *ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations*. Available online at: https://www.intestcom.org/page/31 (accessed August 7, 2019).

Kettler, R. J., Elliott, S. N., Beddow, P. A., and Kurz, A. (2018). "Accessible instruction and testing today," in *Handbook of Accessible Instruction and Testing Practices,* eds S. N. Elliot, R. J. Kettler, A. P. Beddow, and A. Kurz (Cham: Springer), 1–16.

Kieffer, M. J., and Lesaux, N. K. (2007). Breaking down words to build meaning: morphology, vocabulary, and reading comprehension in the urban classroom. *Read. Teacher* 61, 134–144. doi: 10.1598/RT.61.2.3

Kieffer, M. J., Lesaux, N. K., Rivera, M., and Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: a meta-analysis on effectiveness and validity. *Rev. Educ. Res.* 79, 1168–1201. doi: 10.3102/0034654309332490

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educ. Rev.* 78, 333–368. doi: 10.17763/haer.78.2.70783570r1111t32

Mislevy, R. J. (2018). *Sociocognitive Foundations of Educational Measurement*. London: Routledge.

Mislevy, R. J., Steinberg, L. S., and Almond, R. G. (2003). On the structure of educational assessment. *Meas. Interdiscip. Res. Perspect.* 1, 3–62. doi: 10.1207/S15366359MEA0101_02

National Research Council (2014). "Literacy for science: Exploring the intersection of the next generation science standards and common core for ELA standards, a workshop summary," in *Steering Committee on Exploring the Overlap between "Literacy in Science" and the Practice of Obtaining, Evaluating, and Communicating Information. Board on Science Education, Division of Behavioral and Social Sciences and Education,* eds H. Rhodes and M. A. Feder, Rapporteurs (Washington, DC: The National Academies Press), 1–114.

Oliveri, M. E., Ercikan, K., and Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Appl. Meas. Educ.* 27, 286–300. doi: 10.1080/08957347.2014.944305

Oliveri, M. E., and von Davier, A. A. (2016). Psychometrics in support of a valid assessment of linguistic minorities: implications for the test and sampling designs. *Int. J. Test.* 16, 205–219. doi: 10.1080/15305058.2015.1099534

O'Reilly, T., Deane, P., and Sabatini, J. (2015). *Building and Sharing Knowledge Key Practice: What do you Know, What Don't You Know, What Did You Learn*? (Research Report RR-15-24). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12074

O'Reilly, T., and Sabatini, J. (2013). *Reading for Understanding: How Performance Moderators and Scenarios Impact Assessment Design* (Research Report RR-13-31). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2013.tb02338.x

O'Sullivan, B., and Weir, C. J. (2011). "Test development and validation," in *Language Testing: Theories and Practices,* ed B. O'Sullivan (Basingstoke: Palgrave Macmillan), 13–32.

Paas, F., Renkl, A., and Sweller, J. (2003). Cognitive load theory and the instructional design: Recent developments. *Educational Psychologist,* 38, 1–4.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2009). *English Gigaword fourth edition LDC2009T13*. Philadelphia, PA: Linguistic Data Consortium. Available online at: https://catalog.ldc.upenn.edu/LDC2009T13

Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., and Ginsburgh, M. (2009). *Guidelines for the Assessment of English Language Learners*. Princeton, NJ: ETS. Available online at: https://www.ets.org/s/about/pdf/ell_guidelines.pdf

Rose, D., and Meyer, A. (2000). Universal design for learning. *J. Special Educ. Technol.* 15, 67–70. doi: 10.1177/016264340001500307

Rose, D. H., and Strangman, N. (2007). Universal design for learning: meeting the challenge of individual learning differences through a neurocognitive perspective. *Univ. Access Inform. Soc.* 5, 381–391. doi: 10.1007/s10209-006-0062-8

Russell, M., Hoffmann, T., and Higgins, J. (2009). Meeting the needs of all students: a universal design approach to computer-based testing. *Innovate* 5:4. Available online at: https://www.learntechlib.org/p/104243/ (accessed August 7, 2019).

Sato, E. (2008). "Part II: a guide to linguistic modification: Increasing English language learner access to academic content," in *Linguistic Modification* (Washington, DC: LEP Partnership), 53–101. Available online at: https://ncela.ed.gov/files/uploads/11/abedi_sato.pdf

Sato, E., Rabinowitz, S., Gallagher, C., and Huang, C.-W. (2010). *Accommodations for English Language Learner Students: The Effectiveness of Linguistic Modification of Math Test Item Sets (NCEE 2009-4079)*. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Sheehan, K. M., Flor, M., Napolitano, D., and Ramineni, C. (2015). *Using TextEvaluator® to Quantify Sources of Linguistic Complexity in Textbooks Targeted at First-Grade Readers Over the Past Half Century* (Research Report No. RR-15-38). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12085

Sheehan, K. M., Kostin, I., Napolitano, D., and Flor, M. (2014). The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *Element. School J.* 115, 184–209. doi: 10.1086/678294

Snyder, K. K. (2016). *The relationship of formative assessment to the professional development and perspective transformation of teachers* (theses). Student Research, and Creative Activity: Department of Teaching, Learning and Teacher Education, 69. Available online at: http://digitalcommons.unl.edu/teachlearnstudent/69 (accessed August 7, 2019).

Spache, G. (1953). A new readability formula for primary-grade reading materials. *Element. School J.* 53, 410–413. doi: 10.1086/458513

Thurlow, M., Lazarus, S. S., Albus, D., and Hodgson, J. (2010). *Computer-Based Testing: Practices and Considerations* (Synthesis Report 78). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Turkan, S., and Lopez, A. A. (2017). "Helping English language learners access the language and content of science through the integration of culturally and linguistically valid assessment practices," in *Teaching Science to English Language Learners,* eds L. C. de Oliveira and K. Campbell Wilcox (Cham: Palgrave Macmillan), 163–190. doi: 10.1007/978-3-319-53594-4_8

Turmo, A., and Elstad, E. (2009). What factors make science test items especially difficult for students from minority groups? *Nordic Stud. Sci. Educ.* 5, 158–170. doi: 10.5617/nordina.348

Vygotsky, L. (1978). *Mind in society: Development of higher psychological processes (M.* Cole, Ed.). Cambridge, UK: Harvard University Press.

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave MacMillan.

Wiliam, D. (2011). *Embedded Formative Assessment*. Bloomington, IN: Solution Tree Press.

Wolf, M. K., Guzman-Orth, D., Lopez, A., Castellano, K., Himelfarb, I., and Tsutagawa, F. S. (2016). Integrating scaffolding strategies into technology-enhanced assessments of English learners: task types and measurement models. *Educ. Assess.* 21, 157–175. doi: 10.1080/10627197.2016.1202107

Young, J. W., King, T. C., Hauck, M. C., Ginsburgh, M., Kotloff, L., Cabrera, J., et al. (2014). *Improving Content Assessment for English Language Learners: Studies of the Linguistic Modification of Test Items* (Research Report No. RR-14-23). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12023

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.