Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 11, June 2022 ISSN 1531-7714

Measuring Student Reading Comprehension Performance: Considerations of Accuracy, Equity, and Engagement by Embedding Comprehension Items within Reading Passages

Meg Guerreiro, *NWEA*
Elizabeth Barker, *NWEA*
Janice Johnson, *NWEA*

Reading comprehension is measured differently between classroom and more formal approaches to assessment. Traditional reading comprehension assessments often prompt students to read a multi-paragraph passage prior to displaying a set of questions that are related to the passage; however, this approach is not utilized during classroom practices. The study suggests that assessments may inadvertently measure extraneous constructs (e.g., working memory, attention, language, reading ability) by prompting students to answer items at the conclusion of the reading passage. The current study evaluates the effect of asking items throughout the passage (i.e., embedding items) to achieve a more precise measure of reading comprehension by removing barriers for students to demonstrate their understanding. Results showed a significant impact of embedding comprehension items within reading passages on the measurement of student achievement in comparison to answering items at the end of the passage. This may be a more valid approach to measurement of reading comprehension resulting in improved student reading comprehension scores. This approach also has the potential to become a more equitable measurement of reading comprehension by removing barriers to measurement, particularly for marginalized groups (e.g., students with disabilities, memory-load difficulties, English language learners, test anxiety).

Literature Review & Problem Statement

For years, researchers and teachers have wanted to understand the best possible way to measure a student's reading ability. This measurement process includes both precise measures as well as the individualization of reading to support research suggesting that students interact with text in different ways (Stanovich, Cunningham, and Feeman, 1984). Traditional reading comprehension assessment facilitates one method of student-item interaction: it pushes a student to read the entirety of a passage, hold the information in memory, and then answer comprehension questions. Embedding items throughout the passage utilizes universal design for learning (UDL) to provide students with multiple ways to answer an item. For example, embedding items into the reading passages gives students the opportunity to interact with an item (i.e., answer a question) immediately after reading a section of text in which the targeted evidence is presented. This change removes barriers of access (e.g., memory load, attention, reading ability) and potentially allows for a more valid measure of reading comprehension. Embedded items allows students to pause and check for understanding at different points during the reading process. The purpose of incorporating this approach is to create an

assessment that lets students drive the interaction based on their individual needs in reading. The prototype accomplishes this by applying universal design principles and allowing for flexibility in student response, which models good instructional principles. This paper aims to use an equity lens to explore the effect of embedded items within reading passages on student experience and performance on reading comprehension assessments.

Problem Statement

Reading comprehension assessment aims to measure student understanding of text. During formative classroom assessment, teachers often measure reading comprehension by pausing and checking for understanding at different points in the passage the students are reading; yet, within formal assessment, reading comprehension is often tested in a very different way. A typical reading comprehension item facilitates just one method of student-item interaction: a passage followed by the presentation of items related to the passage content. This layout often prompts students to read the entire passage prior to answering the questions; thereby, extraneous variables are introduced (e.g., working memory, language, reading level) that may be barriers for students to demonstrate their true reading comprehension of portions of the text. By designing an item that allows for students to pause and check for understanding at specific locations in the passage, an assessment can model best practices of teaching and formative assessment of reading comprehension by assessing reading comprehension after students have just read the portion of text in which the evidence is presented (Fisher & Frey, 2015). This approach aims to help reduce anxiety, reduce memory load (especially with language acquisition), better align with student reading preference, and support students with disabilities while still supporting high-quality, text-centered, standards-based assessment.

Theoretical Framework

The current research uses item design founded in evidence-based practices such as UDL (CAST, 2018), scaffolding, segmentation of text (Abedi et al., 2010), think-aloud self-regulation techniques (Carioli & Peru, 2016), and vocabulary in context. Through implementation of these approaches, the current study aims to provide a more valid measurement of reading comprehension—as measured by student achievement—while also mitigating barriers of access for students, particularly students with disabilities. This approach aligns closely with Vygotsky's (1978) zone of proximal development, which suggests major child development occurs when the child collaborates with an adult and, over time, creates mastery and independence. This framework helps to create the foundation for scaffolding that has been defined as a "process that enabled a child or novice to solve a problem, carry out a task or achieve a goal which would be beyond his unassisted efforts" (Wood et al., 1976, p. 90). The current study builds on this foundation by providing a scaffolded approach to the assessment of reading comprehension while implementing UDL principles and other foundational approaches to establish a level of independence for the student when interacting with and understanding text.

Universal Design for Learning. According to the Higher Education Opportunity Act of 2008:

> 'universal design for learning' means a scientifically valid framework for guiding educational practice that—(a) provides flexibility in the ways information is presented, in the ways students respond or demonstrate knowledge and skills, and in the ways students are engaged; and (b) reduces barriers in instruction, provides appropriate accommodations, supports, and challenges, and maintains high achievement expectations for all students, including students with disabilities and students who are limited English proficient. (p. 122 STAT.3088).

As such, the core of UDL is comprised of three principles: multiple means of engagement, multiple means of representation, and multiple means of action and expression (Meyer & Rose, 1998). Creating a classroom that provides multiple means of engagement, representation, and action provides a more inclusive environment for all students and ways to interact with content to meet individual needs, including the needs of those with disabilities. UDL puts the what, how, and why of learning at the center of instructional planning, and this incorporation allows students to experience UDL features prior to a formal assessment (Roski, Walkowiak, & Nehring, 2021). For example, materials made with simplified, clear language will support all students, not just students with reading difficulties and emergent bilinguals. The connection of a UDL classroom to a UDL assessment allows for a

more authentic experience for the student.

Within assessment, universally designed items allow for the greatest range of student access and tend to result in a more valid measurement of performance of all students (Johnstone et al., 2008). When an assessment is created with UDL in mind, barriers (e.g. background knowledge, language, technology knowledge) are removed—or never introduced—to reduce construct-irrelevant variance, resulting in more accurate results of student growth and achievement. These UDL principles should be incorporated in the construction of the items, as well as item layout design, and the inclusion of accessibility features (Johnstone et al., 2008). The UDL foundation ensures access into the content and allows assistive technology to function properly, providing the opportunity for most students to engage with the content. For example, a UDL-based assessment ensures a student with a visual impairment or blindness can use a screen reader and refreshable braille device to navigate and access the reading material; however, UDL does not stop there. The design of the item is just as important including the layout of the text, syntax and grammatical structure, and flexibility of how a student responds all need consideration (McKenna et al., 2013).

Substantial research has been conducted to understand the effects of test accommodations provided for students, specifically the use of UDL in an attempt to build an assessment environment to support the needs of all students. Assessments that use UDL have the potential to create measurement that supports all students. The UDL approach is particularly beneficial for students with disabilities because it removes barriers that lead to construct-irrelevant variance affecting performance. Construct-irrelevant variance refers to aspects of a task that are either too easy or too difficult for reasons unconnected with the skill or content the item intends to target; these aspects lead to scores—higher or lower—that do not accurately represent an affected students' knowledge of the targeted skill or content (Messick, 1995). For example, an assessment that aims to measure adding two-digit numbers may also require a student to read extensive text to access the mathematical construct. In this case, the additional presence of the construct of reading may interfere with how an early reader makes sense of and responds to the mathematical construct and negatively impact the construct's measurement. The foundational approach of UDL aims to provide the student a more equitable approach to testing prior to offering more individualized accommodation.

Despite the foundational benefits of UDL within assessment, many assessments of reading comprehension fall short in modeling this approach. Within a formative assessment context, specifically in reading comprehension, UDL allows for students to engage with contextual understanding during the reading process (Abedi et al., 2011), but formal reading assessments are not designed in this way.

Scaffolding. Learning to read is a process that changes over time, with readers moving from learning about print, to decoding, to fluently making meaning from text (Brown, 1999). These changes require individualized instructional supports to promote continued student growth (Brown, 1999). A scaffolding approach occurs naturally within classroom instruction and has been overwhelmingly deemed as essential and vital to comprehension (Clark & Graves, 2004; Duffy, 2002). Scaffolding allows for specific texts to be used at various times in a reader's development (Brown, 1999). Scaffolding is commonly seen in adaptive testing, in which item difficulty is based on students' current level of performance; as students progress to more difficult skills, the texts presented to the students require more independence because of the demands they place on student characteristics such as fluency and background knowledge (Brown, 1999). This allows assessments to move away from a one-size-fits-all approach.

The scaffolded approach can also be modeled for items within assessments, providing adaptivity to items based on student knowledge. This is also commonly seen in adaptive testing. Taking scaffolding one step further allows for specific item display at the point in which the targeted portion is presented for the student to read. This provides an additional form of scaffolding, allowing for the student to interact deeply with a portion of text in the moment rather than at the end of the text. This is similar to the method used in textbooks to help students learn content and has shown to produce gains for struggling students while underscoring the need for individualized supports (Callender & McDaniel, 2007). Similarly, hypermedia annotations in texts, which allow students to interact more deeply with content during the reading process, have shown significant benefits to passage

comprehension and vocabulary (Abuseileek, 2011), suggesting continued scaffolding to support students within reading comprehension.

Segmenting Text. Literature shows that students with disabilities perform at substantially lower levels than students without disabilities (33% proficient to 75% proficient) (Abedi et al., 2010). This difference may be attributed to disability, but other factors may also be present, such as frustration, fatigue, or lack of testing accommodations (Abedi et al., 2010). One accommodation used for students with disabilities is segmenting text into meaningful units (Abedi et al., 2010). The use of this technique suggests that organizing text into smaller units may facilitate recall and improve comprehension for some readers (Abedi et al., 2010), allowing them to digest the information they have read (Abedi et al., 2010) and to demonstrate comprehension more accurately. In a study conducted by Abedi et al. (2010), researchers focused on the need to shift assessment design to better support students with disabilities by allowing for built-in breaks to maximize students' working memory and to reduce the potential for disengagement. This segmented approach to assessment proved to be not only a more reliable measure of student abilities to accurately comprehend text but also a more accessible way for students with disabilities to engage with the content (Abedi et al., 2010).

Think Aloud. The think-aloud strategy is another approach used within reading comprehension that allows students to ask themselves questions to ensure that they comprehend what they have read before they move on (Fisher et al., 2011; Carioli & Peru, 2016). Block and Israel (2004) recognized that thinking aloud provided the ability for a student to use their metacognitive skills, a powerful tool for deeper reading comprehension. This approach—pausing and checking for understanding during the reading experience—can help students make meaning of text during the reading process; it can provide valuable data to teachers on students' development of mental models and comprehension of text. Although students are not literally talking aloud, they are engaging in scaffolded checks for understanding, which encourages the use of metacognitive skills to make meaning of text. As shown by Fisher et al. (2011), when teachers model think-aloud strategies, students' ability to engage in this metacognition increases, resulting in increased student achievement—specifically, adding to student self-assessment of their own comprehension.

Vocabulary in Context. Vocabulary knowledge plays a major role in reading comprehension and is often targeted in large-scale assessments because of the ease of both developing and scoring items (Qian, 2008). For example, assessments will typically ask students to use the content of the text to help define a specific word within the passage. Context plays an important role in this approach to measuring reading comprehension (Qian, 2005). Qian (2008) argues the importance of contextualized vocabulary within assessment, specifically for emergent bilinguals, encourages both understanding and learning vocabulary words in context. Researchers suggest contextualized vocabulary allows for more real-world application and models actual communication patterns (Qian, 2008).

Within a formal assessment context, vocabulary is typically assessed both in stand-alone items and in association with a set of passage-based items. However, within traditional reading comprehension assessments, items that measure vocabulary appear at the end of the text, along with other non-vocabulary items. This removes the targeted word from its contextual point in the passage and requires students to use prior knowledge of the word, draw on their memory of the context, or skim the passage to find the original location and reread the context. This traditional approach encourages students to answer quickly, using potentially faulty memory of the context, or puts pressure on students' ability to locate a single word within an entire paragraph or passage of text.

Current Embedded Item Design

The approach in the current study enhances the layout of reading comprehension assessments by embedding items alongside the reading passage. This approach supports the effort to remove physical, cognitive, and sensory barriers that are unrelated to the measurement of the students' reading comprehension, thereby mitigating the effect of construct-irrelevant variance. Embedding items in reading passages gives students the opportunity to interact with an item immediately after reading a section of text in which the targeted evidence is presented. The goal of this approach is to minimize barriers and allow for flexible functionality. The approach also aims to improve the performance of students, specifically students with disabilities, and to reduce memory load, anxiety, and

physical barriers (Abedi et al., 2010; Johnstone et al., 2006), thereby providing more reliable measurement results in comparison to less flexible comprehension item designs that often rely on the student reading the entire passage before answering any items. Consequently, the embedded item is designed to allow for variation of student access, input, and output of information (e.g., flexibility in how and when students respond to items within the passage), so students can select the approach to answering questions that is best suited for their needs without compromising the accuracy and measurement of the assessment (Johnstone et al., 2006; Meyer et al., 2014). This flexibility increases inclusivity to various subpopulations, minimizes bias, and allows for better accommodation support. This is supported by UDL assessment research (Thompson et al., 2002), which emphasizes a framework that is flexible in the ways students respond, reduces barriers, and maintains high expectations.

Within assessment design, the incorporation of features such as scaffolding, text segmentation, think-aloud strategies, and vocabulary in context is intended to support a more valid and reliable measure of reading comprehension while also adhering to the UDL framework by providing multiple ways for students to interact with the item. While the current embedded-item design was founded on the concepts of scaffolding, think-aloud strategies, and text segmentation, items were also designed to resemble a more natural approach to reading, thereby allowing students to interact with the text in various ways and to check for understanding at the point of the passage that best meets their individual needs (Lapp, Fisher, & Grant, 2008).

The embedded-items approach to measuring reading comprehension offers many potential benefits. Although the design centers the needs of students with disabilities, the benefits of the embedded-item design extend beyond one population, using a UDL approach to assessment that ensures access for many (e.g., emergent bilinguals, talented and gifted, students with anxiety, students with attention difficulties, struggling readers). Through the UDL approach and item flexibility, students have the option to stop and check for comprehension at the point in the passage when the evidence appears or at the end of reading the entire passage. This flexibility support students' individual reading preferences, while reducing construct-irrelevant variance.

A main construct that the embedded-items approach aims to mitigate is the measurement of working memory, a fundamental skill lacking in beginning readers (Dehaene, 2009) and therefore limiting the ability for beginning readers or emergent bilinguals to interact with items. Traditional assessment that places questions at the conclusion of the passage or that include the directive that students should read the entire passage before responding to the items assessing understanding of smaller segments of text may be drawing on student memory and attention instead of measuring true understanding of the information from the passage. While some items are better suited to be answered after the student has read the entire passage, such as those covering whole-text concepts including main idea or author's perspective, other items may be tightly connected to a single sentence or paragraph. By integrating reading comprehension items throughout the passage, the measurement results are more closely aligned with students' true reading comprehension and may not be as reliant on additional irrelevant constructs such as memory or attention.

Content: Texts and Items

Assessment approaches must be tailored to the content they target to provide students with an opportunity to show what they are capable of achieving. When considering the embedded-items approach to reading assessment, assessment creators must carefully evaluate embedded items against the critical content-related components of high-quality reading assessment: texts and items.

Reading is an activity dependent on the presence of text; to read, one must have something to read. Therefore, the selection of text is of vital importance when assessing reading (CCSS Initiative, 2010). In addition to adhering to the foundational requirements for any assessment around sensitivity, fairness, and lack of bias laid out in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014), the selected texts should represent writing that is published or of publishable quality, should include factually accurate information, and should provide a sense of completeness even if excerpted. Furthermore,

following the guidance in the Common Core State Standards (CCSS) and the supplemental information published later that incorporated new research, the Supplemental Information for Appendix A, the text should be of suitable complexity for the grade level of the students being assessed (Sheehan, 2017). That complexity should be determined using multiple quantitative measures and a thorough qualitative analysis that considers complexity across the dimensions of levels of meaning/purpose, structure, language convention and clarity, and knowledge demands (CCSS, 2010).

Any rigorous assessment of reading must be built on a foundation of text that displays all the aforementioned characteristics, so it is worth considering whether the embedded-items approach will allow for use of this type of text. In fact, the embedded-items approach works well with high-quality, grade-level, complex text. First, use of embedded items encourages the use of extended text rather than the use of short excerpts that fail to fully develop the ideas and features common in authentic reading experiences (Carver, 1994). Second, because embedded-item design supports access to rich text by students who have disabilities or who are thought to read below grade level, there is less pressure to include short or overly simplistic texts to seemingly meet the needs of those students. As a side benefit, assessment practices do drive instruction, despite efforts to the contrary (Connor, 2019), and student performance on embedded item types signals that all students can access rich grade-level text and models approaches to support students in their reading and analysis. While the security surrounding high-stakes assessment environments prevents teachers from directly leveraging test events and test content as teachable moments for practicing reading strategies or for supplying individualized scaffolding that stretches a student's zone of proximal development, it is to be hoped that aligning these assessments to instructional best practices makes student experiences of tests more positive and makes any teaching-to-the-test that occurs before or afterward more productive of actual gains in reading comprehension.

Once a text is selected, reading assessment must seek to understand if a student comprehends the text. It is evident that students show comprehension in a myriad of ways. For example, students demonstrate comprehension of a set of directions when they can play a game. They also show comprehension through the act of reaching eagerly for the sequel of a book to find out what happens next, through the reenactment of an exciting story on the playground, or through the use of information from an article to support their point in a classroom discussion. In formal assessments, the central question of comprehension is answered through carefully structured items (Reed & Kershaw-Herrera, 2015; Abedi et al., 2011). The question of comprehension cannot be answered through items that target superficial or trivial information, nor through items that merely require matching the wording in the item to details in the text. Instead, to draw a valid conclusion about a student's comprehension of a text, items must address the major ideas and literary elements present in the text and the aspects of craft that are central to its meaning (CCSS Initiative, 2010). The analyses students are required to perform must be grounded in the text and be developmentally appropriate. In formal assessment, this can be achieved through items that address the intent and detail of grade-level standards such as the CCSS or other college- and career-ready standards.

The embedded-items approach lends itself well to assessing the CCSS, as shown in Table 1. Table 1 identifies examples of literary and informational reading standards for which embedded items could be particularly effective because they require a focus on specific portions or locations within the text (CCSS, 2010).

While not listed in Table 1, vocabulary standards are also particularly suitable for assessment via location-specific embedded items that focus students' attention on areas of the text. A traditional assessment item might excerpt the sentence in which a target word or phrase occurs, as shown in Figure 1. The necessary context for students to use in figuring out the meaning of that target word is rarely present within that single sentence except at the lowest grades, requiring students to track back into the passage as a whole. Even with a parenthetical paragraph citation, as shown in Figure 1, there is effort involved in skimming the paragraph to find the actual occurrence of the word. Setting an embedded item adjacent to the occurrence of the target word assists students in leveraging the context beyond that sentence and seeing the word as it relates to the wider text. However, care must be taken that if the

Table 1. CCSS Evidence Supporting the Use of Embedded Items for Grades 2–5 (CCSS, 2010).

Grade Level	Standard Code	Standard Text
Grade 2	CCSS.ELA-LITERACY.RI.2.2	Identify the main topic of a multi-paragraph text as well as the focus of specific paragraphs within the text.
Grade 3	CCSS.ELA-LITERACY.RI.3.8	Describe the logical connection between particular sentences and paragraphs in a text (e.g., comparison, cause/effect, first/second/third in a sequence).
	CCSS.ELA-LITERACY.RL.3.5	Refer to parts of stories, dramas, and poems when writing or speaking about a text, using terms such as chapter, scene, and stanza; describe how each successive part builds on earlier sections.
Grade 4	CCSS.ELA-LITERACY.RI.4.5	Describe the overall structure (e.g., chronology, comparison, cause/effect, problem/solution) of events, ideas, concepts, or information in a text or part of a text.
	CCSS.ELA-LITERACY.RI.4.8	Explain how an author uses reasons and evidence to support particular points in a text.
Grade 5	CCSS.ELA-LITERACY.RI.5.8	Explain how an author uses reasons and evidence to support particular points in a text, identifying which reasons and evidence support which point(s).

Note. (CCSS, 2010).

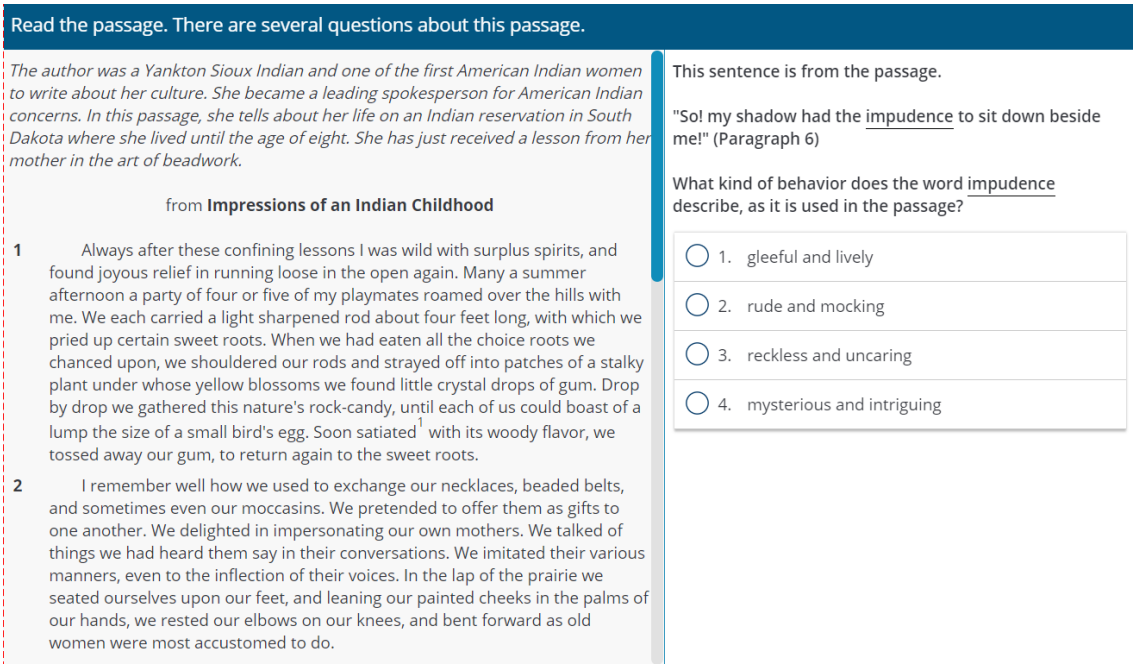
context necessary for understanding a target word follows it at a significant remove – perhaps in subsequent paragraphs –this word may not be suitable that sentence and seeing the word as it relates to the wider text. However, care must be taken that if the context necessary for understanding a target word follows it at a significant remove – perhaps in subsequent paragraphs –this word may not be suitable for an embedded approach; students would not be able to respond accurately without reading further and an embedded item gives tacit approval to pausing at that point. If this is the case, a vocabulary item may be more properly considered a 'whole-text item' and presented at the end of the passage alongside items about the text's main idea.

Evaluation of the way standards intersect with embedded items is not limited to its applicability to individual standards or even entire strands like Vocabulary. The CCSS laid out three key shifts for English Language Arts instruction, curriculum, and assessment that continue to serve as touchstones and should be considered with the embedded-items framework.

The first key shift of the CCSS calls for "regular practice with complex texts and their academic language" (CCSS, 2010, para. 2). While instruction and curriculum are implicated via the reference to 'regular practice,' assessment must also confront the demands of this key shift by ensuring that students are assessed on their comprehension of progressively more complex texts, texts that include academic language (CCSS, 2010). As discussed above in reference to the selection of texts for assessment, embedded items support assessment with complex texts. Furthermore, embedded items provide an efficient means of targeting vocabulary, encouraging direct assessment of the academic language referenced in this key shift.

The second key shift is "reading, writing, and speaking grounded in evidence from texts, both literary and informational" (CCSS, 2010, para. 6). One of the great problems for reading assessment over the years has been the tendency to go beyond the text, asking for inferences that are not actually warranted by the information in the text or asking questions that vaguely reference generalizations about the text. Writing embedded items forces test developers to indicate the

Figure 1. Traditional item example showing how a sentence could use an excerpt from the text demonstrating the misalignment between the item and the passage prompting students to scroll, seek, and find the point in the passage where the evidence appears.



location of items, tying them to specific portions of the text or explicitly placing them at the end of the text in acknowledgement that they are whole-text items. Items that are related the entirety of the text, or passage, for standards that address localized information will be red flags for poor alignment given that students do not need to read the passage in its entirety in order to effectively respond to the item. Moments when test developers debate the correct location for embedded items will prompt reflection on whether the item is directly grounded in text evidence.

The third key shift is "building knowledge through content-rich nonfiction" (CCSS, 2010, para. 9). Again, this is very clearly a signal for the arenas of instruction and curriculum but also has significance for assessment in that assessment texts should include a helping of content-rich nonfiction (Connor, 2019). The embedded-items approach will aid students in successfully comprehending information-dense nonfiction, as well as in demonstrating that comprehension. This can create a positive feedback loop as student success on assessments pushes for more opportunities to practice in those successful areas.

Current Study Overview

Prototype Development Process

To study the impact of embedded-item layout on student reading comprehension performance, the study included a mixed-methods user experience based quasi-experimental within-subjects randomized control design. This process allowed the use of a research-based approach to examine item usability, layout, and design through a foundation of user experience testing. As a result, a final prototype was developed and data were collected from students through both the control assessment (traditional item layout) and the experimental assessment (embedded-item layout). Quantitative analyses were used to assess the effects of assessment type on student reading comprehension performance.

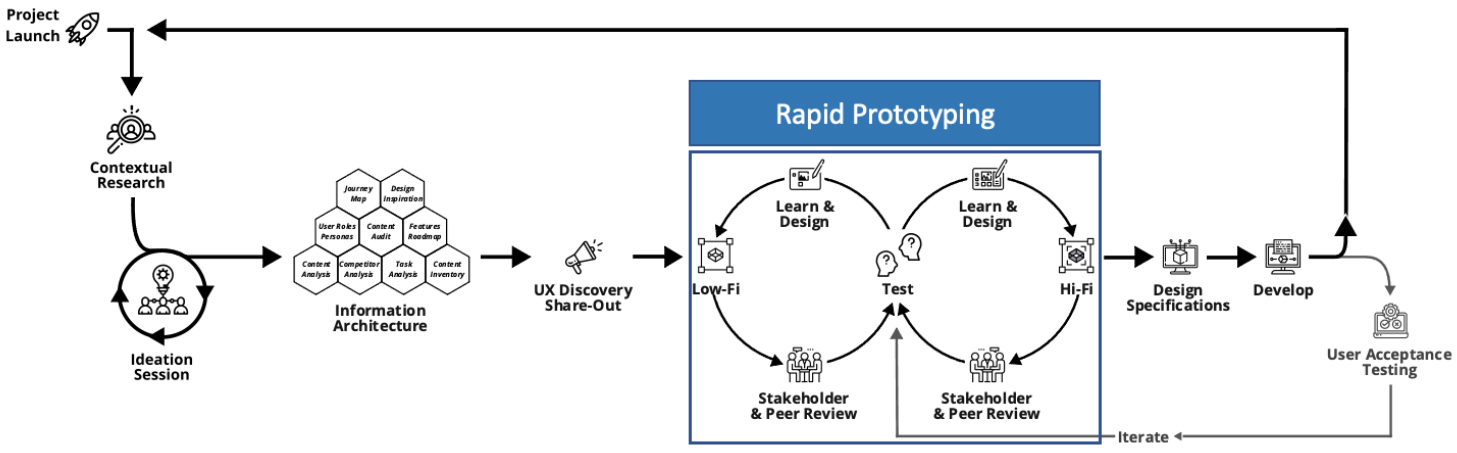
The prototype development process was a collaborative effort and iterative process bridging reading content experts, UDL experts, developers, researchers, and user experience designers with the goal of receiving stakeholder input, complete user testing, and create a working assessment item for deeper data collection, see Figure 2. The goal of this

phase was to build a functional prototype. Development of the prototype began with identifying the assessment purpose, theory of action, and evidentiary argument. Specific accessibility features to include were also identified such as speech to text, accessible to assistive technology (e.g., screen reader and refreshable braille device), accessible fonts and colors, easily identifiable symbols, intuitive layout, and minimal scrolling. The prototype began through contextual research and ideation sessions to scope an initial design and approach. Developers and user experience designers were brought in to help scope the design and determine technical aspects of development. Then, a wireframe was created to present the basic components and layout with the goal to obtain expert feedback. Changes were made and the first version of a working prototype was created.

Each prototype included one passage with five associated multiple-choice items that measured reading comprehension for Grade 3. The items were developed by content experts and aligned to the CCSS (see Table 1). The difficulty of items was equivalent between all prototypes. The items associated with each passage covered a range of skills appropriate to embedding within the passage as well as allowing assessment of comprehension of the passage as a whole. Each set of five items included two context-clues vocabulary items that were well-suited to the embedded approach as well as a third item targeting understanding of a specific paragraph of each passage. The remaining two items for each set assessed whole-text comprehension via questions about main idea, organizational structure, or author's purpose.

This balance of the embedded-items approach targeting small segments and whole-text items assessing overall comprehension used in this prototype is important to note; assessment using this intentional balance of embedded items and whole-text items means students must still grapple with a passage's big ideas and must still form an accurate mental model of the text as a whole. This remains a critical component of measuring reading comprehension. However, students can also easily and accurately leverage their comprehension of specific segments of text to respond to embedded items as part of their process in arriving at that overall model. Content experts selected passages and associated items based on their comparability in terms of passage genre and structure, text complexity, item targets, and item difficulty. Then, internal iterations were implemented to finalize the item layout for usability testing. Emphasis was placed on user experience design to (a) explore a possible better approach to reading comprehension; (b) meet the student where they are academically; and (c) evaluate the impact to the total student experience. The user center design approach (focusing on users and unique needs) implemented a design-based foundation, analysis of information architecture and user experience, iterations on the design including stakeholder and peer review, and development of design specifications and final working prototype.

Figure 2. User-centered design approach to prototype development



Usability Testing Process

Throughout 2016–2018, two phases of usability and field test studies were conducted on the final working item prototype. Since the item type and design were new, the goal of usability testing was to ensure intuitiveness, ease of use, students' opinion on the approach, and overall user experience. Through this approach, a variety of data were collected including user experience, student progressions, and qualitative interviews. Usability testing Phase I included students (n = 10) in grades 2–6 and user experience data were collected including data on intuitiveness of design, layout and symbols, favorability of model and discuss any possible confusion with this approach. The grade span was purposeful to ensure that new and early readers were included in the usability testing with the assumption that if new readers were able to access the items, proficient readers would likely be able to access the items. Data was used to understand the student experience across various demographics including design, layout, or text that may have caused confusion. Results from usability testing in Phase I included general student confusion of symbols and colors/patterns used throughout the item. Phase I item revisions aimed to mitigate this highlighted confusion.

Usability testing in Phase II included updated styles, directions, and layouts that were reviewed by a second subset of students (n = 10) in grades 2–6, similar to the makeup of the sample from Phase I but included a different sample of students. Another round of user experience data were collected on final design updates. This component of the study focused more on usability design and research for product development to draft more comprehensive research questions for the small-scale pilot phase of the work. Data was used to understand the student experience ensuring the changes made from Phase I were appropriate and the layout, design, and text was intuitive to navigate. Results from Phase II included a redesign of the item layout and location of components, improvements to the item tutorial, and general student confusion of updated symbols and colors/patterns used throughout the item. Similar to Phase I, Phase II item revisions aimed to mitigate overall confusion.

Field Testing

Usability testing was followed by field testing in Phase III that implemented the input from Phases I and II and evaluated student performance across item types (i.e., embedded items versus traditional common stimulus items). The review of research supports the need for further investigation of approaches to design a UDL assessment experience as well as to support better measurement of reading comprehension. The research question that guided the field testing and main component of this study aimed to explore what relationship is found between student outcome performance (as measured by overall score) and assessment item type (embedded or traditional common stimulus items) with the hypothesis that there would be a difference in student performance between assessment item type. The findings from this research could help inform policy makers and approaches to assessment design, particularly as focus shifts to more authentic and equitable assessment practices. This work also aims to contribute to the conversation on how we can improve our measurement of reading through common stimulus item types; an approach that may indicate a removal of barriers for measurement and demonstration of knowledge could help provide a more valid assessment and outcome as well as better and more actionable data for teachers.

Field testing used a within-subject counter-balanced quasi-experimental design measuring reading comprehension performance on both assessment types. Data were analyzed using a paired samples t-test. The final item type used in the Phase III field testing consisted of a Grade 3 common stimulus reading passage (eight paragraphs) with five multiple-choice items. The item layout also included two ways to access the items: a center column of items aligned with the passage and a top navigation bar of items. The item also included a section in which the item is displayed. A screenshot of the embedded item type is displayed in Figure 3, in comparison to the traditional item design in Figure 4. Three multiple-choice items were designed to be embedded with the option of answering them during the point of the passage in which the evidence appeared. These items (items 1–3) are shown in the column directly to the right of the passage and are aligned mid-passage in text. Two multiple-choice items were related to the entire passage, thereby appearing at the end of the passage. These items (items 4–5), similar to the embedded items, are shown in the column directly to the right of the passage but appear at the bottom (i.e., at the end of the passage). All items (mid-passage and whole text) only displayed in the item

presenter if they were selected. Once an answer was selected, a green bar appeared under the item number indicating the item had been answered.

Given the foundational goal of implementing a UDL approach to this item layout, the multiple item access panels (vertical column and horizontal top row) allowed for students to interact with the items and passage whichever way they felt more comfortable, thereby creating a true UDL approach to an assessment of reading comprehension. As a result of this flexibility (i.e., students can switch between answering items as they read the passage or answering items traditionally after reading the passage in its entirety), items are a fixed set and can only be adapted at the passage level.

Methods

Study Design

The assessments were administered in the school computer lab and included two passages, each with five associated multiple-choice items: (a) new embedded layout where items appear mid-passage and (b) traditional layout where items appear at the end of the passage. The design included a within-subjects randomized control design where the embedded layout served as the experimental item and the traditional layout served as the control item. Students were randomly assigned to a starting passage (either embedded or traditional) followed by the second item type (i.e., a student who received the traditional item

Figure 3. Embedded items screenshot showing three embedded items and two thematic items. Layout also depicts universal design option (column of items) and traditional approach (question toolbar).

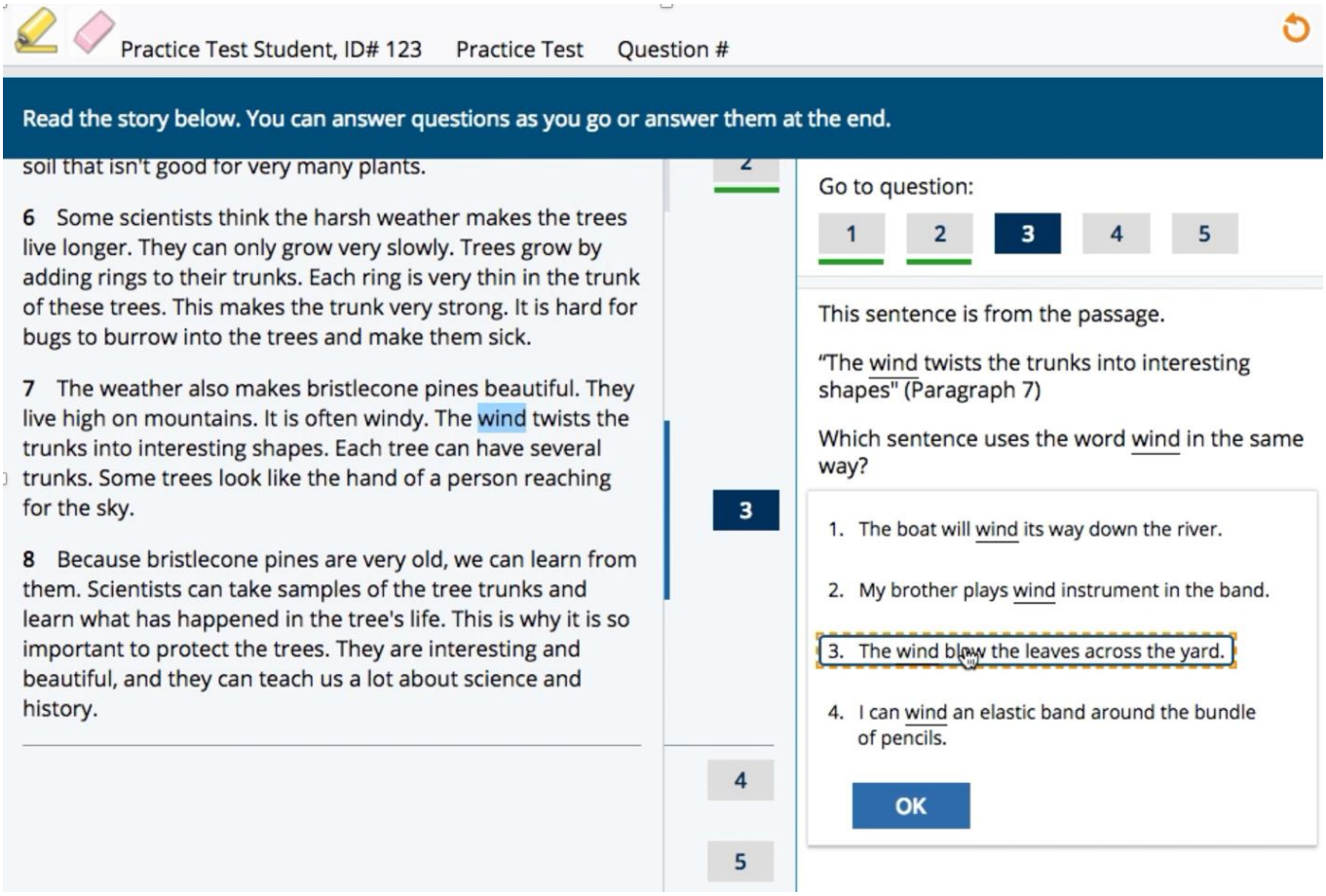
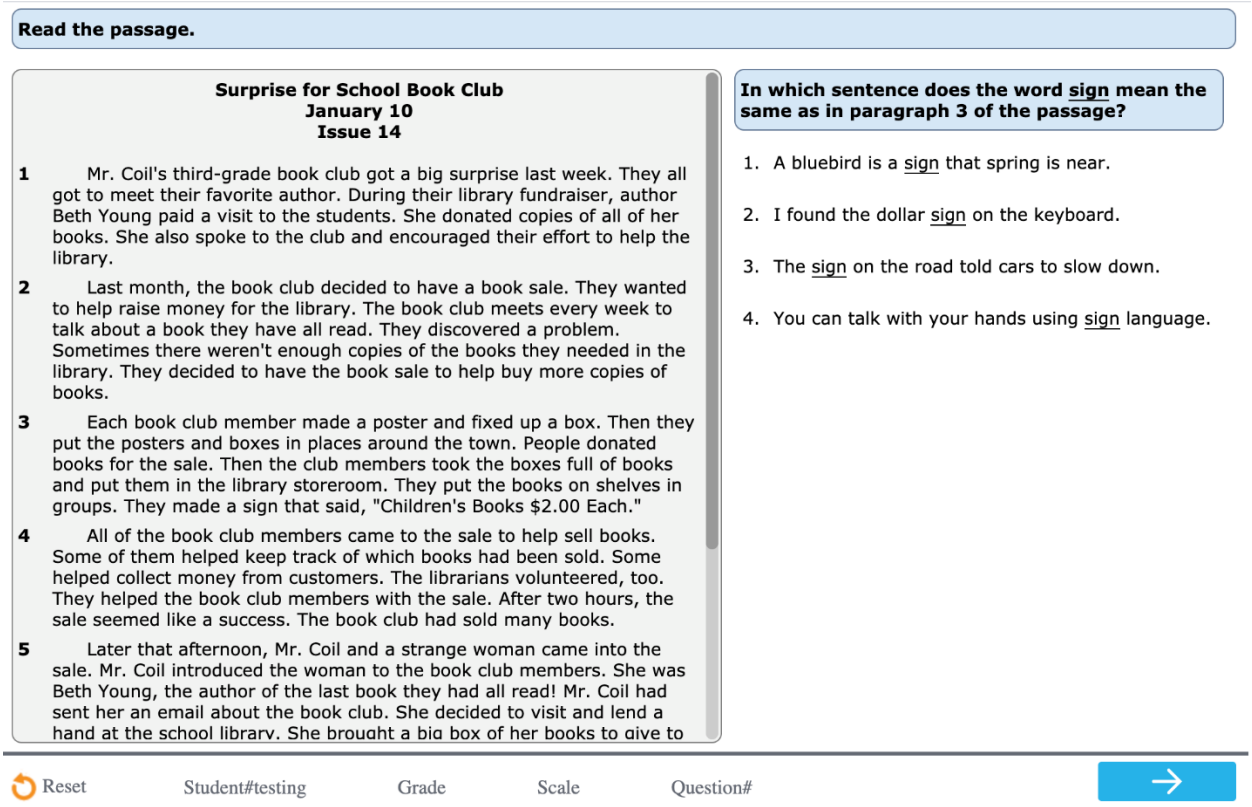


Figure 4. Traditional item design screenshot showing the reading passage on the left with questions displayed on the right, not in line with text.



layout first would receive embedded-items layout second, and vice versa), and student performance on both item types were measured. Performance was measured using a RIT score with a range from 100 to 350. "RIT scores relate directly to the RIT vertical scale, an equal-interval scale that is continuous across grades" (NWEA, 2019, p. 53).

Participants

The participants included a convenience sample (n = 130) of Grade 3 students from one K–12 Midwestern public elementary school. The sample included 37.7% students who identified as male, 38.5% who identified as female, and 16.9% unidentified. Race/ethnicity of the sample included 38.5% identifying as White and 61.5% identifying as non-White. The sample also included both ELLs (18.5%) and students with disabilities (11.5%). Demographic data for the sample are displayed in Table 2. Performance was measured using two five-item common stimulus reading comprehension passages (i.e., one using the traditional item layout and another using the embedded-item layout). See Figures 3 and 4. The difficulty level of both the traditional and embedded layout were equivalent. See Tables 4 and 5.

Analysis

Phase I and Phase II used qualitative user experience observation and interview data to improve item design and layout. Data collection analysis centered on student experience with the item type, and analysis explored the platform architecture to ensure that student response data were collected appropriately. A design review also ensured that students understood how to navigate a new item type. Additional internal iterations helped to provide insight into item design and layout. Phase II resulted in final design specifications and working prototype. Phase III used quantitative data from both the traditional and embedded layouts. Assessment results were reviewed using descriptive statistics to gain insight into patterns and comparisons. Data were further analyzed using an independent means paired samples t-test to assess the differences in student reading comprehension

performance between item types (traditional and embedded).

Table 2. Demographic Data (n = 130)

Variable	*n* (%)
Female	59 (45.4)
Male	49 (37.7)
Unidentified	22 (16.9)
White	50 (38.5)
Non-White	80 (61.5)
ELL	24 (18.5)
TAG	18 (13.8)
SPED	15 (11.5)
504	5 (3.8)
FRL	17 (13.1)
RTI	30 (23.1)

Note. *n*(%)

Results

Phases I and II

Qualitative data in the form of observation and interviews were used to provide item feedback and accessibility. Main changes from Phases I and II included designation of location and type of symbol used to denote an item embedded within the text, layout of passage and items, intuitiveness of item experience, ease of directions, and overall engagement with this new experience. Results were used to create the final item prototype used in Phase III and displayed in Figure 3.

Phase III

Data showed a significant increase in student performance between traditional common stimulus items and embedded items (see Table 3). On average, students scored higher on the embedded-item assessment (M = 227.91, SE = 3.28) compared to the traditional item assessment (M = 189.43, SE = 0.76). This difference, 38.48, 95% CI [32.30, 44.66], was significant $t(114) = 12.33$, $p = .000$, and represented a medium-sized effect, $r = 0.57$. This is also displayed in Figure 5 showing students performed significantly better on the embedded items in comparison to traditional items; 49 students (38% of students) scored a perfect (5 out of 5) or nearly perfect (4 out of 5) in comparison to the traditional item layout where three students (2%) scored nearly perfect (4 out of 5) and zero students scored perfect (5 out of 5). Additional information about the RIT scores for items correct between traditional and embedded items used in these analyses are displayed in Tables 4 and 5. These results suggest that an embedded items-formatted assessment provides a more accurate representation of student achievement; students demonstrated higher achievement when completing passaged-based reading assessment answering items as they read rather than at the end. It is worth noting that these higher results do not indicate the assessment was easier in an artificial way; rather, with the removal of barriers (e.g. working memory, attention, anxiety) this allows for a more valid measure. In fact, RIT scores (see Tables 4-5) demonstrate the substantial difference in difficulty between the items in the traditional passage and the items in the embedded passage, with a perfect score on embedded items 87 RIT points higher (more difficult) than a perfect score on the traditional passage.

Discussion

The study aimed to explore what relationship is found between student outcome performance (as measured by overall score) and assessment item type (embedded or traditional common stimulus items). The new embedded items type was designed with UDL considerations to create a more equitable and valid approach to the measurement of reading comprehension with common stimulus items. A mixed-methods user experience-based quasi-experimental within-subjects randomized control design was implemented to investigate the effect of item type on student outcome. Findings from the study suggest that students perform significantly better on an assessment using an embedded-items format in comparison with an assessment using a traditional items format. These results suggest that utilizing an embedded-items format to measure reading comprehension (i.e., presenting questions to students throughout the reading experience instead of at the end of a passage) may provide better, and more valid, outcome scores for students. This does not indicate an easier assessment; rather, a more valid assessment due

to the potential removal of barriers (e.g. working memory, attention, anxiety) for demonstration of understanding. Results from an embedded-items assessment aim to increase validity as well as provide better and more actionable data for teachers. These findings intend to create a foundation for future exploration in improving measures of reading as well as provide an example of an equitable and universally-designed assessment. Future work should explore psychometric considerations to validate assumptions as well as include a more robust sample to explore effects of item type based on student demographic characteristics.

Limitations

This study is primarily limited by the nature of the schools in which it was conducted and is not generalizable beyond the sample. Usability testing was conducted with a small sample of students (n = 20), which may not be generalizable or account for other ways students interact with items. Analyses do not explore difference between various subgroups. Methodologically, the analysis does not account for the nesting of data between classrooms. The items used for analysis were considered retired items and may have content limitations, deeming them unusable for true academic achievement measures. Since the embedded item type was a new format, additional directions were provided to students prior to beginning the assessment to help them understand how to use the embedded-items feature. However, researchers are unable to fully ascertain if students used the embedded-items feature or if they answered embedded items using the traditional approach. Despite this limitation, researchers used observation and student conversations to infer that most students did use the embedded-items approach.

Implications for Practice

For years, researchers have tried to understand the best way to measure student reading ability including both precision of measurement and individualized supports (Stanovich, 2984). Yet, despite this push for better measures of reading ability and the recent push for a UDL framework within education, few assessments meet the criteria, specifically within English Language Arts. A traditional reading comprehension item facilitates one method of student-item interaction, which pushes a student to read the entirety of a passage, hold the information in memory, then answer comprehension questions even if some of the items query only parts of the mental model the student has formed while reading rather than the model as a whole. This approach may result in

Table 3. Paired Samples t-test (n = 115)

	M	SD	SE mean	*t*-value	*df*	*p*
Traditional	189.43	8.20	0.76	12.33	114	.000
Embedded	227.91	35.17	3.28			

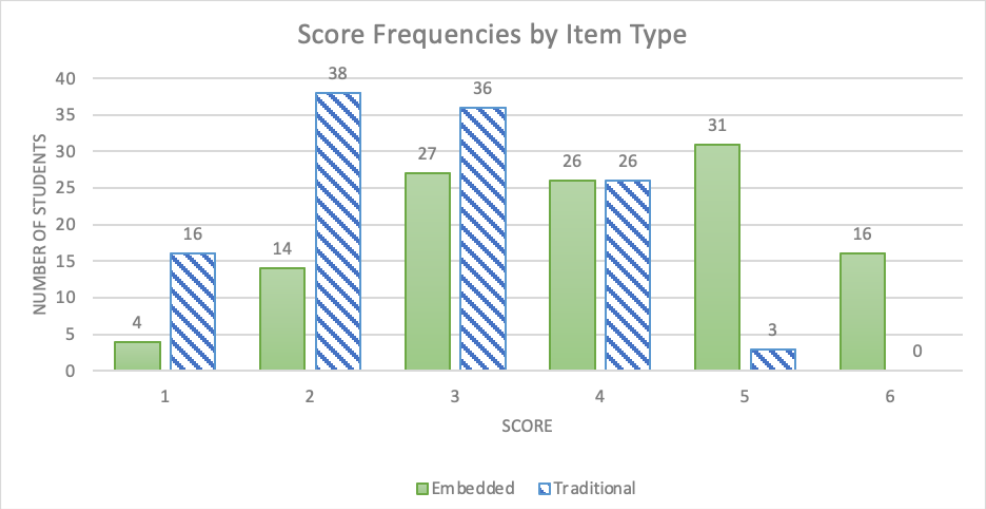
Table 4. Score Frequencies for Embedded Items (n = 125)

Number correct	RIT	*n* (%)
0	133.15	4 (3.2)
1	193.81	14 (11.2)
2	209.23	27 (21.6)
3	221.97	31 (24.8)
4	237.39	33 (26.4)
5	298.04	16 (12.8)

Table 5. Score Frequencies for Traditional Items (n = 119)

Number Correct	RIT	*n* (%)
0	180.72	16 (13.4)
1	181.60	38 (31.9)
2	191.65	36 (30.3)
3	199.95	26 (21.8)
4	210.00	3 (2.5)
5	210.87	0 (0)

Figure 5. Graph displaying student performance on embedded items (solid green) and traditional items (striped blue).



inaccurate performance outcomes or disengagement, specifically for marginalized groups. The purpose of this study was to create a new item type that responds to different student needs in reading using the frameworks of UDL, segmentation of text, and think-aloud structure to support most students. The embedded item type also aim to remove barriers that could potentially affect reading outcomes such as minimizing the need for the student to search for answers throughout the passage or holding information in short-term memory. These improvements have the ability to minimize the effect of assessment time, working memory, attention difficulties, and struggles with language structure. Results show that students performed significantly better when answering embedded items within a text. This may be a more valid approach to measurement of reading comprehension resulting in improved student scores along with a possible increase in student engagement. This approach also has the potential to become a more equitable measurement of reading comprehension by removing barriers to measurement, particularly for marginalized groups (e.g., students with disabilities, memory-load difficulties, ELLs, test anxiety). It is also important to note that some students may select not to engage with items during the reading process. The current study allowed for this flexibility (i.e. universal design approach). This is something that should be explored in the future.

In order to create measures of reading that are more equitable with the inclusion of richer data, it is critical to reexamine traditional approaches and it is important that future studies fully validate this approach to measuring reading comprehension. Future areas of exploration should aim to include purposeful items scaffolded within the passage to better understand students' development of mental models and text comprehension throughout the reading

process. This approach could elicit richer data driven from more purposeful content. It would also be meaningful to examine how student choice, more culturally-responsive and authentic text, and sufficient background knowledge impact outcomes. Within the scope of the current study, future research is needed to explore achievement and engagement outcomes by specific demographic groups, with specific attention devoted to exploring how emergent bilingual students respond to the embedded-items approach.

References

Abedi, J., Kao, J. C., Leon, S., Mastergeorge, A. M., Sullivan, L., Herman, J., & Pope, R. (2010). Accessibility of Segmented Reading Comprehension Passages for Students with Disabilities. *Applied Measurement in Education, 23*(2), 168–186. https://doi.org/10.1080/08957341003673823

Abedi, J., Leon, S., Kao, J., Bayley, R., Ewers, N., Herman, J., & Mundhenk, K. (2011). *Accessible reading assessments for students with disabilities: The role of cognitive, grammatical, lexical, and textual/visual features* (CRESST Report 784). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abuseileek, A. F. (2011). Hypermedia annotation presentation: The effect of location and type on the EFL learners' achievement in reading comprehension and vocabulary acquisition. *Computers and Education, 57*(1), 1281–1291. https://doi.org/10.1016/j.compedu.2011.01.011

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

Block, C.C., & Israel, S.E. (2004). The ABCs of performing highly effective think alouds. *The Reading Teacher, 58(2),* 154-167.

Brown, K. J. (1999). What kind of text - For whom and when? Textual scaffolding for beginning readers. *Reading Teacher, 53*(4), 292–307. https://doi.org/10.2307/20204794

Callender, A. A., & McDaniel, M. A. (2007). The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology, 99*(2), 339–348. https://doi.org/10.1037/0022-0663.99.2.339

Carver, R.P. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: implications of instruction. *Journal of Reading Behavior, 26*, 413-437. https://doi.org/10.1080/10862969409547861

Carioli, S. & Peru, A.(2016). The think-aloud approach: a promising tool for online reading comprehension. *Journal of Media Literacy Education, 8(1),* 49-61.

CAST (2018). Universal Design for Learning Guidelines version 2.2. Retrieved from http://udlguidelines.cast.org

Common Core State Standards (CCSS) Initiative (2010, June). *Common core state standards for English language arts and literacy in history/social studies, science and technical subjects*. Washington, DC: Council of Chief State School Officers & National Governors Association.

Connor, C. M. (2019). Using technology and assessment to personalize instruction: preventing reading problems. *Prevention Science, 20*(1), 89–99. https://doi.org10.1007/s11121-017-0842-9

Clark, K. F., & Graves, M. F. (2004). Scaffolding students' comprehension of text. *International Reading Association*, 570–580.

Dehaene, Stanislas. (2009). *Reading in the brain: The new science of how we read.* Penguin Books.

Duffy, G. G. (2002). The case for direct explanation of strategies. In C.C. Block & M. Pressley (Eds.), *Comprehension instruction: Research-based best practices* (pp. 28–31). New York: Guilford.

Fisher, D., & Frey, N. (2015). Checking for understanding digitally and during content area learning. *The Reading Teacher: Rethinking Literacy Instruction, 69(3)*, 281-286.

Fisher, D., Frey, N., & Lapp, D.(2011). Coaching middle-level teachers to think aloud improves comprehension instruction and student reading achievement. *The Teacher Educator, 46,* 231-243. https://doi.org10.1080/08878730.2011.580043

Johnstone, C., Altman, J., Thurlow, M., & Moore, M. (2006). Universal design online manual. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. https://nceo.info/Resources/publications/UDmanual

Johnstone, C.J., Thompson, S.J., Bottsford-Miller, N.A., & Thurlow, M.L. (2008). Universal design and multimethod approaches to item review. *Educational Measurement: Issues and Practices, 27(1),* 25-36.

Lapp, D., Fisher, D. & Grant, M. (2008). "You can read this text – I'll show you how": interactive comprehension instruction. *Journal of Adolescent & Adult Literacy, 51*(5), 372–383.

McKenna, M.C., Labbo, L.D., Kieffer, R.D., Reinking, D. (2013). International handbook for literacy and technology. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Messick, S. (1995). Validity of psychological assessment validation of inferences from persons' responses and performances in scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Meyer, A. & Rose, D. H. (1998). Learning to read in the computer age. Cambridge, MA: Brookline Books.

Meyer, A., Rose, D.H., & Gordon, D. (2014). Universal design for learning: theory & practice. Wakefield, MA: CAST Professional Publishing.

National Center for Education Statistics. (2016). The condition of education.

NWEA (2019). Technical Manual for MAP Growth. Portland, Oregon.

Reed, D. K., & Kershaw-Herrera, S. (2015). An examination of text complexity as characterized by readability and cohesion. The Journal of Experimental Education, 84(1), 75–97. https://doi.org/10.1080/00220973.2014.963214

Roski, Malte, & Nehring. (2021). Universal Design for Learning: The More, the Better? Education Sciences, 11(164), 164. https://doi.org/10.3390/educsci11040164

Sheehan, K. M. (2017). Validating automated measures of text complexity. *Educational Measurement: Issues and Practice, 36*(4), 35–43.

Stanovich, Cunningham, A. E., & Feeman, D. J. (1984). Intelligence, Cognitive Skills, and Early Reading Progress. Reading Research Quarterly, 19(3), 278–303. https://doi.org/10.2307/747822

Thompson, S., J., Johnstone, C.J., & Thurlow, M.L. (2002). Universal design applied to large scale assessments (synthesis report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/Synth44.pdf

Qian, D. D. (2005). Demystifying lexical inferencing: The role of aspects of vocabulary knowledge. *TESL Canada Journal, 22*(2), 34–54.

Qian, David D. (2008). From single words to passages: Contextual effects on predictive power of vocabulary measures for assessing reading performance. *Language Assessment Quarterly, 5*(1), 1–19. https://doi.org/10.1080/15434300701776138

Roski, M., Walkowiak, M., Nehring, A. Universal Design for Learning: The More, the Better? Educ. Sci. 2021, 11, 164. https://doi.org/10.3390/educsci11040164

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological process*. Cambridge, MA: Harvard University Press.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*, 89–100.

Citation:

Guerreiro, M., Barker, E., & Johnson, J. (2022). Measuring Student Reading Comprehension Performance: Considerations of Accuracy, Equity, and Engagement by Embedding Comprehension Items within Reading Passages. *Practical Assessment, Research & Evaluation*, 27(11). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/11/

Corresponding Author:

Meg Guerreiro
NWEA

Email: meg.guerreiro [at] nwea.org