



Performance of Technology-Enhanced Items in Grades 1–12 English Language Proficiency Assessments

Prepared by
Ahyoung Alicia Kim
Rurik L. Tywoniw
Mark Chapman



**Wisconsin Center for
Education Research**
SCHOOL OF EDUCATION
UNIVERSITY OF WISCONSIN-MADISON

Contents

Executive Summary.....	3
1. Introduction	5
Research Questions	7
2. Literature Review	7
Item Types in Reading Assessments	8
Defining and Categorizing Technology-Enhanced Items	10
Advantages and Limitations of TEIs	12
<i>Psychometric properties of TEIs.</i>	14
<i>Efficiency of TEIs.</i>	16
<i>Accessibility of TEIs.</i>	17
3. Methods.....	19
Study Context.....	19
Data.....	25
<i>Student test data.</i>	25
<i>Telemetry data.</i>	27
<i>Item Response Theory (IRT) analyses.</i>	27
<i>Item efficiency.</i>	29
<i>Accessibility and universal tools.</i>	30
<i>Item design analysis.</i>	30
4. Findings	31
Findings on research question 1: Item Characteristics	31
Findings on research question 2: Efficiency of TEIs vs. MCIs.....	41
<i>Item duration of TEIs and MCIs.</i>	41
<i>Efficiency of TEIs and MCIs.</i>	44
Findings on research question 3: Accessibility of TEIs vs. MCIs.....	47
5. Discussion.....	51
6. Conclusion.....	56
References	59

Executive Summary

Technology-enhanced items (TEIs) are innovative, computer-delivered test items that require test takers to interact with the test environment in ways beyond those required when responding to traditional multiple-choice items (MCIs). The interactive nature of TEIs allows them to potentially measure aspects of the test construct better than MCIs (Sireci & Zenisky, 2006); however, despite the popularity of TEIs in computer-based assessments, there is little research that compares students' performance on TEIs to their performance on MCIs in English language proficiency (ELP) assessments. In addition, there is little understanding of how TEI innovations affect accessibility for multilingual learners. Previous research on TEIs is limited to math and science domains (Crabtree, 2016), and research of TEIs in ELP contexts is rare, especially in kindergarten to grade 12 (K–12) settings.

The purpose of this study is to examine the performance of grades 1–12 English learners (ELs) on TEIs and MCIs in an online reading test. The test is part of ACCESS for ELLs, an annual large-scale ELP assessment administered to over 2 million K–12 ELs in the U.S., and consists of four domains (Listening, Reading, Speaking, and Writing). We analyzed 1.2 million ELs' scores on the Reading domain test of ACCESS Online across five different grade-level clusters: 1, 2–3, 4–5, 6–8, and 9–12. The test included content-matched TEIs and MCIs; that is, pairs that shared the same content but differed in response mode. Content-matched TEIs and MCIs were evaluated for standard item performance metrics such as difficulty, discrimination, and information using Item Response Theory (IRT) modeling. In addition, item efficiency was investigated using the amount of information provided in relation to item duration. Finally, to examine how TEIs affect the accessibility of the test, ELs' use of several online universal tools (e.g., magnifier, highlighter) was tracked and correlated with each item type.

Overall, TEIs proved slightly more difficult than content-matched MCIs, but they did not differ in discriminative power. The information provided by TEIs to the overall test varied by grade level, with the TEIs typically being more informative for ELs at higher grade or proficiency levels. TEIs generally had longer item durations than their MCI counterparts, yet TEIs were on average more efficient than MCIs in grades 6–8, providing more information for these select grades. Furthermore, TEIs elicited more use of universal tools, especially the highlighter and line-guide tools, across all test takers. These quantitative results, along with a qualitative analysis of reading item design features, provide insights that can guide further development of TEIs for online ELP assessments for multilingual learners, allowing test developers to embrace the interactivity of TEIs while mitigating potential issues.

1. Introduction

With the advancement of language testing technology in computerized environments, online assessments have incorporated potentially more authentic and engaging item types in the form of technology-enhanced items (TEIs). The TEIs differ from traditional multiple-choice items (MCIs) in terms of how students engage with the test interface and select a response. For example, *hotspot* items allow test takers to click parts of a larger response area, and *drag-and-drop* items require test takers to move images to a certain location within a larger picture.

The development of TEIs reflects the educational field's desire to create items that reflect the overall construct measured on the assessments (Sireci & Zenisky, 2006). However, development and implementation of TEIs does not guarantee benefits to assessment authenticity, reliability, and validity compared to traditional MCIs or constructed-response items (Huff & Sireci, 2001; Russell, 2016). Potential benefits of TEIs include the broadening of measured constructs, presentation of authentic contexts allowing test takers to better demonstrate their knowledge, and reduced guessing. Yet our understanding of these benefits is limited by the lack of psychometric data on these item types vis-à-vis traditional item types, as well as the difficulty of generalizing from the existing data due to the variability in how different item types perform (Parshall & Harmes, 2014; Sireci & Zenisky, 2006). According to Bryant (2017), findings on TEIs are particularly confounding, especially regarding their introduction of construct-irrelevant variance alongside a potential increase in test taker engagement and motivation. There is thus a need for evidence-based arguments to understand the influence of technology enhancements on assessments.

However, despite ongoing research of TEIs in educational assessments (Crabtree, 2016; Dolan et al, 2011; Gutierrez, 2009; Huff & Sireci, 2001; Jodoin, 2003; Thomas, 2016; Wan &

Henly, 2012; Woo, Kim, & Qian, 2014), especially in science, technology, engineering, and mathematics (STEM) fields, few empirical studies on the performance of TEIs exist in the field of language assessment, particularly of young learners, and no studies have compared TEIs to traditional items that are paired in content, differing only in interaction features. Previous studies have compared TEI and traditional items which differed in multiple ways, such as format of information presentation, inclusion of media, knowledge and content targets, and scoring procedures. Thus, it is difficult to pinpoint exactly what about TEIs may affect item performance on educational assessments. This study seeks to address this gap by evaluating TEIs that differed from traditional items only in the response interaction, with the topic domains, targeted knowledge, point value, and scoring procedures between the TEIs and the traditional items remaining the same. Three overarching aspects of TEIs are evaluated alongside traditional MCIs, each of which has been examined in previous research: psychometric properties such as difficulty and discrimination (Crabtree, 2016; Gutierrez, 2009), efficiency (Jodoin, 2003; Wan & Henly, 2012), and accessibility (Kim et al, 2019; Russell, 2016).

The TEIs examined in this study are part of the Listening and Reading domain tests of ACCESS for ELLs¹ (hereafter ACCESS). ACCESS is a high-stakes standardized English language proficiency (ELP) assessment that is widely used in the K–12 context. It was developed by WIDA in collaboration with the Center for Applied Linguistics, and it is annually administered to over 2 million K–12 English learners (ELs) across 40 U.S. states and territories. ACCESS measures the academic ELP development in the four language domains—Listening, Reading, Speaking, and Writing. ACCESS is offered in both paper-and-pencil and online

¹ ACCESS for ELLs is offered in both paper-and-pencil format (ACCESS Paper) and online format (ACCESS Online). This study focuses on the online format of the assessment, which incorporates technology-enhanced items. Kindergarten ACCESS for ELLs is administered in paper only; therefore, it is not included in this study.

formats. Although the TEIs included in ACCESS Online are expected to perform similarly to traditional MCIs and fit the same measurement model, little is known regarding how students' engagement with and performance on TEIs differs from traditional MCIs, suggesting the need for more research on this topic.

The purpose of this study was to explore the inclusion of TEIs in the reading domain test of ACCESS Online. It compares multiple aspects of students' performance on TEIs vs. MCIs, which were paired in terms of content (see Methods section for details). In detail, the study addresses TEI psychometric qualities and performance equivalence using Item Response Theory (IRT) analytic methods, and further investigates efficiency and accessibility vis-à-vis more traditional MCI counterparts. Item response data were analyzed considering both students' grade and proficiency levels. Moreover, test design characteristics were qualitatively analyzed through sequential explanatory design (Creswell & Plano Clark, 2018) to examine the design differences between select content-matched TEI and MCI pairs. Using this mixed-methods approach, the following research questions were addressed in the study.

Research Questions

1. How do TEIs compare to traditional MCIs in terms of difficulty, discrimination, and information?
2. How do TEIs compare to traditional MCIs in terms of efficiency?
3. How do TEIs compare to traditional MCIs in terms of test takers' use of universal tools?

2. Literature Review

Before beginning our own investigation of TEIs on ACCESS Online, we present a review of research on various issues related to the use of TEIs. This begins with a discussion of how different types of items are employed in reading assessments of English learners. A narrower

discussion of technology-enhanced item types follows, with a detailed review of research examining specific properties of TEIs in a multitude of assessments. We conclude this section with a grounded purpose of the current study.

Item Types in Reading Assessments

In language testing, reading tests often consist of traditional MCIs that include a reading passage, a prompt, and three or four answer options. However, these item types, typically presented in written texts, may lack authenticity compared to the real-life reading tasks that they are intended to emulate (Alderson, 2000; Bennet, 1999). Language assessment researchers have examined the effect of item format on the reading construct (Currie & Chiramanee, 2010; In'nami & Koizumi, 2009; Katalayi & Sivasubramaniam, 2013; Lim, 2019; Pae, 2018) and suggest using a wide range of item types on second language (L2) reading tests. Overreliance on a single item type may restrict the ability to test the intended range of language skills, an idea which problematizes the prevalence of traditional stem-and-options multiple-choice items in L2 reading tests. Rupp, Ferne, and Choi (2006) examined the reading section of a standardized language proficiency test that used only traditional multiple-choice items, specifically analyzing test-taking strategies elicited from the think-alouds of 10 university-level language learners at a Canadian university. They found that multiple-choice items elicit a specific reading construct related to problem-solving rather than comprehension, which has implications for the validity of reading tests that use only traditional multiple-choice items.

However, complete elimination of the traditional MCI is not desirable, considering the efficiency of the item type in measuring test takers' reading ability. Rather, the goal is to diversify the reading comprehension test format by adding different item types, including TEIs. When including other reading item types, in addition to MCIs, test designers must understand

that even subtle adjustments to item formats may cause differences in item performance and examinee cognition. Moon, Keehner, and Katz (2019) examined different selected response formats, each renditions of the true-false format, on a mathematics test. Findings indicate that test taker performance and time taken to answer were affected by small item-type distinctions such as the presence of “do not know from given information” answer choices and the ability to select multiple options. It is thus important, especially in high-stakes large-scale standardized testing, to critically evaluate the alternatives to traditional MCIs afforded by computer-based testing (CBT).

Although language testers are interested in alternatives to the traditional MCI format, the discussion of TEIs in the CBT context has mostly existed in STEM fields. There has been some move to CBT in first language (L1) assessment or measuring English language arts (Scalise, 2012), but literature on using TEIs in L2 and ELP testing is still limited (Rasskazova et al, 2017). Despite much discussion of the use of computers as a teaching and testing medium for ELs (Chapelle & Douglas 2006; Douglas & Hegelheimer, 2007; Jamieson, 2005), little empirical investigation exists regarding how TEIs are employed in ELP testing.

For example, Papadima-Sophocleous (2008) described the use of an adaptive computer-based test for university-level L2 English placement which used traditional MCIs alongside TEIs, such as items with visual prompts and fill-in-the-blank items with drop-down menus. She found mixed results regarding the difficulty and discrimination of the test, indicating that some technology-enhanced reading comprehension items performed poorly in discriminating between high and low achievers. However, no TEIs and traditional MCIs were paired for comparison of their effects on test performance. The current study thus hopes to push the field forward by examining the use of MCIs vs. TEIs in a large-scale standardized ELP assessment. Before

describing the current study further, a more detailed description of the current state of TEIs is warranted.

Defining and Categorizing Technology-Enhanced Items

TEIs can refer to a wide range of items, as reflected in the variety of definitions available in the field. Parshall, Harmes, Davey, and Pashley (2010) consider TEIs to be test items that use technologies or computer features to deliver assessments that cannot be easily presented in traditional paper-and-pencil format. Smarter Balanced, developer of K–12 content assessments across multiple states in the U.S., defines TEIs more concretely as “computer-delivered items that include specialized interactions for collecting response data. These include interactions and responses beyond traditional selected-response and constructed-response” (Smarter Balanced Assessment Consortium, 2012, p. 1). Similarly, Bryant (2017) describes TEIs as “computer-based items that make use of formats and/or response actions not associated with conventional MCs and constructed-response items” (p. 2). By drawing upon these definitions, this study defines TEIs as items delivered via computer-based assessments that use methods or media for item stimuli or responses that are unavailable in paper formats and different from traditional multiple-choice or constructed-response items.

There are multiple ways to categorize TEIs. Russell (2016) offers two general categorizations of TEIs: *technology-enabled items*, which may function similarly to traditional items but use media and content only available in a computer-based test, and *true technology-enhanced items*, which involve response interaction spaces native to a computerized test environment. This categorization also acknowledges that TEIs can fit into both categories.

Scalise and Gifford (2009) focus on response types and categorize TEIs according to the extent the items constrain test takers’ responses. These include (1) fully constrained (i.e.,

multiple choice-like), (2) selection/identification, (3) reordering/rearranging, (4) substitution/correction, (5) completion, (6) construction (e.g., short answers, essays, and figure construction), and (7) presentation (e.g., performance assessments, possibly involving peer-interaction). They emphasize how TEIs provide item design possibilities which are more practically scorable than paper-and-pencil constructed response items, while still allowing for more input and interaction from the examinee than traditional multiple-choice items.

While Scalise and Gifford (2009) categorize TEIs according to the item response type, Parshall et al. (2010) categorize TEIs based on seven dimensions: (1) *assessment structure*, such as item type and how responses are collected (e.g., constructed-response item), (2) *response action* required by test takers to complete the item, such as drag and drop; (3) *media inclusion*, such as the use of audio or visuals in prompt or options; (4) *level of interactivity* in terms of how an item reacts to examinee input (e.g., images updating based on examinee choices); (5) *complexity* of the distinct objects and pieces of information test takers must process and interact with; (6) *fidelity* of the degree to which the item resembles authentic real-world contexts; and (7) *scoring model*, which refers to how response data are collected and scored.

Existing TEI categorizations discussed above focus on the format, design, and response actions of items. What is missing from the above categorizations is information on the type of knowledge, skills, or abilities these different categories of TEIs elicit from examinees, indicating a need for more research in this regard (Schechinger, 2012). In addition to understanding TEIs and how items identified by that term might be categorized, it is necessary to examine the potential benefits and shortcomings of using such items to have a comprehensive understanding of them (Parshall et al., 2010).

Advantages and Limitations of TEIs

The use of TEIs in CBT offers both potential benefits and drawbacks. Research on TEIs has compared the construct equivalence of TEIs and their more traditional counterparts. Early studies conducted on innovative computer-based items on the Graduate Record Exam (GRE) found that items designed specifically for the digital platform performed similarly to MCIs while being based on a distinct construct (Bennet et al., 1999; Bennet & Sebrechts, 1997). Wan and Henly (2012) compared the construct equivalence of two constructed-response TEI types with traditional MCIs on science achievement tests across multiple grade levels (grades 5, 8, and high school) using confirmatory factor analysis. They found that a one-factor model (indication of a single underlying construct) accurately explained performance variance among students. These findings suggest that TEIs do not necessarily introduce construct-irrelevant variance, which bolsters the case for including them in assessments.

Beyond equivalence, TEIs are also of interest for their potential to enhance tests and gather test taker responses in a way that resembles authentic, real-world tasks, thereby adding fidelity to item response actions (Bennet & Rock, 1995; Parshall, 1999; Parshall, Davey, & Pashley, 2000). This authenticity has the potential to make TEIs more interactive and engaging to learners (Bryant, 2017; Scalise & Gifford, 2006). TEIs offer item writers the ability to draw on templates and response types which are authentic with respect to the target domain, require more interaction from test takers, and engage learners in a way that is natural to computer use (Huff & Sireci, 2001). TEIs can also be designed such that the response process removes construct-irrelevant variance (Messick, 1989) associated with guessing, superficial answer plausibility, and overlap across item stimuli, stem, and options. For example, when images are used for response

options rather than sentences, it may eliminate linguistic cues contained in the correct answer sentence which are not related to comprehension.

At the same time, even TEIs authentic to the intended domain may introduce construct-irrelevant variance if they require computer-use proficiency irrelevant to the construct. Items that require computer proficiency, as in those that require test takers to interact with the computer in a different way than they might interact with a comparable paper-and-pencil test, could introduce characteristics that are irrelevant to the construct being measured on the test (Huff & Sireci, 2001; Sireci & Zenisky, 2006). Thus, test developers must ensure not only that TEIs are comparable in performance to traditional items, but also that TEIs add value rather than construct-irrelevant variance.

Research indicates that certain types of TEIs offer more construct fidelity than others. Russell and Moncaleano (2019) analyzed the content and response action authenticity of 236 TEIs, including 102 English language arts drag-and-drop and select-text items, across large-scale multi-subject standardized academic tests. They found that select-text items were of moderate or high construct fidelity, whereas drag-and-drop items had low fidelity more often, meaning the response process is either inappropriate for the assessed construct or not reflective of real-world interaction with text information. This again underscores the need for research that can more directly compare technology enhancements when TEIs are designed with construct equivalence to traditional items in mind.

Further empirical research on TEIs has focused on the psychometric performance of TEIs (difficulty and discrimination). Understanding that innovative item presentation may distract and slow down test takers, researchers have also compared TEIs and MCIs in terms of item duration. In the case that TEIs take longer to complete than MCIs, researchers are additionally interested

in whether TEIs contribute information to the overall assessment more efficiently than MCIs. In addition, Russell (2016) posits that for TEIs to work effectively, TEIs must be evaluated for more than construct issues. TEIs must also be examined for *usability*, which relates to how a TEI is implemented on the test in terms of functionality and visual layout, and *accessibility*, which relates to how well a TEI allows test takers to access item content. This characteristic may be particularly relevant for English learners or learners with disabilities who might need additional support in accessing test content. The following sections outline research that examined the performance of TEIs in terms of their standard psychometric properties (difficulty, discrimination, and information) in addition to efficiency and accessibility.

Psychometric properties of TEIs. Test developers have been concerned with how well TEIs reliably correspond to overall test scores and how they perform compared to traditional items in terms of difficulty and discrimination. Research suggests TEIs are comparable to traditional items in terms of discriminative power but perform differently across grade and proficiency levels, with some studies finding higher TEI discrimination with low-performing test takers (Gutierrez, 2009) and others indicating better TEI discrimination with high-performers (Crabtree, 2016).

To explore item difficulty and the reliability of TEIs, Jodoin (2003) examined MCIs and TEIs, including drag-and-drop and create-a-tree items (items in which test takers establish relationships among concepts) on a computer programmer certification exam using IRT information measurement. Metrics derived from IRT provide information about how difficult and discriminating an item is, additionally giving test developers a sense of how reliably a test item performs across different ability levels. This level of item performance across test takers' abilities is termed "information" in IRT terminology. In Jodoin's (2003) study, there were a total

of 73 MCIs and 25 innovative items analyzed. In this study, the TEIs were found to be slightly more difficult than MCIs but also more informative (i.e., they contributed more per item to the reliability of the test) than MCIs in general.

Wan and Henly (2012) also compared MCIs with figural response TEIs (i.e., items that test takers respond to by manipulating images, either freely or within some constraint) and constructed-response TEIs on science achievement tests. Findings from IRT analyses indicate that TEIs were typically more informative than MCIs across grade levels (grades 5, 8, and high school), yet the tradeoff between relative gain in informativeness and ease of scoring (scoring was more difficult for more informative constructed-response items) varied between grade levels. These results suggest that TEIs may be more suitable than MCIs for certain grade levels only, as discussed above.

Similarly, Masters and Gushta (2018) compared various item types on a geometry test administered to 425 fourth graders. The item types included constrained, selected-response items ($k = 10$); both traditional button-press MCIs and more interactive MCIs with clickable objects; and less constrained response type TEIs ($k = 10$), such as hotspot, figure placement, matrix completion, and categorization items. Results showed that the innovative TEIs were more difficult, yet also more reliable and informative to the overall test. Moreover the TEIs correlated more highly with constructed-response items which directly assessed knowledge of the geometry construct.

Although the above research suggests TEIs to be more difficult than MCIs, Thomas (2016) did not find such consistent evidence. Thomas (2016) examined over 2,000 public school students taking a computer-mediated math test comprised of 59 tasks which were either multiple-choice, drag-and-drop, or simulation items. The author found that TEIs were not necessarily

more difficult than MCIs at every grade level. This stands in contrast to findings from other researchers, so it remains to be seen if technology enhancement distinctions or item content differences have a larger impact on item difficulty in language assessments.

The research discussed above generally indicates that TEIs are more difficult but can be more informative than MCIs. However, it is unclear whether these findings are generalizable across different age groups and performance levels. Additionally, in the above studies, technology enhancements could not be isolated as the underlying factor in different item performances, as these studies examined characteristics of open-ended TEIs as compared to traditional items. Further research on the performance of TEIs is critical before generalizing those findings. In addition, data from multiple grade levels and minimally enhanced TEIs is paramount to our understanding of the impact of technological enhancements on item properties.

Efficiency of TEIs. Effective TEIs must exhibit satisfactory efficiency. TEIs are intended to provide benefit in the form of innovative means of presenting stimuli, media, and answer responses, but these innovative presentations may introduce unintended complexity. Russell (2016) states that more authentic and interactive TEIs may also come with a time burden and lower test efficiency. If a TEI has a confusing layout or unintuitive interface, or if the innovative features of item affect its functionality, it can lead to the test being less usable and less efficient (Russell, 2016).

The time burden of TEIs has been investigated previously. Partly due to the increased amount of information included, TEIs usually require more testing time than traditional MCIs (Jodoin, 2003; Qian, Woo, & Kim, 2017; Wan & Henly, 2012). However, an increase in the average amount of time a test taker needs to respond to each item (hereafter item duration) is neither a positive nor negative influence on the test if items contribute differing levels of

information to the overall test score. In addition to examining item duration, *item efficiency*, or the ratio of information an item contributes in relation to its average duration (item information divided by duration), has also been considered to help test developers understand TEI usability in educational measurement (Jodoin, 2003; Crabtree, 2016; Wen & Henly, 2012). However, item efficiency has yet to be explored in computer-based language testing.

Wan and Henly (2012) examined the efficiency of TEIs used in a K–12 computer-based science test. Using IRT modeling, they measured item efficiency by dividing item information with average duration needed to respond to the item. They found that TEIs were more informative but less efficient at lower grades, though the item efficiency of TEIs matched that of traditional MCIs at higher grades. Similarly, Qian, Woo, and Kim (2017) compared MCIs against various types of TEIs, which included fill-in-the-blank calculation questions, multiple-response items, and ordered-response items. They found that TEIs, especially fill-in-the-blank items, provided more information to a test score (i.e., are more reliable and discriminating), but required more time to complete, thus lowering the efficiency of those items.

Accessibility of TEIs. TEIs should be designed with accessibility in mind. Accessibility allows all test takers to access the construct of the test by making items meet the needs of a broad range of students (Willner & Monroe, 2016). Thus, TEIs should maximize inclusivity to allow all students to demonstrate their knowledge and abilities. In other words, TEIs should be just as accessible as the traditional item formats that test takers are familiar with. Accessibility is a primary concern when developing assessments and applying principles of universal design (Center for Applied Special Technology [CAST], 2011; Thurlow et al, 2010). Universal design is concerned with providing learners with multiple means of information representation, multiple means of action and expression of knowledge, and multiple means of engagement (see CAST,

2011 for more information). When applied to educational assessment, universal design can involve providing clear language, tools to support understanding, multimedia (text, sound, images, or a combination of these), and variations in response processes.

Universal design can also be applied through the provision of appropriate accessibility features. Accessibility features, also known as universal tools, refers to functionality within a test that can help make the test more comprehensible, legible, bias-free, clear, and straightforward. Color modification overlays, pop-up dictionaries, spell checkers, highlighters, and magnifier functions are all examples of universal tools. Previous research suggests that online test platforms could improve access to testing for ELs (Thurlow, Lazarus, Albus, & Hodgson, 2010) through wider inclusion and better promotion of universal tools. For instance, Kim, Yumsek, Chapman, and Cook (2019) examined the use of universal tools by K–12 ELs taking an ELP computer-based test, finding that universal tools, as intended, were used more frequently on average by students with disabilities. However, no known existing research has examined the effect of TEIs on learners' performance on ELP assessments. The universal tools embedded in assessments could support the accessibility of TEIs (Bryant, 2017), and test takers might use these tools more when completing TEIs as opposed to MCIs. However, media presentation and interactivity in items can provide new challenges for examinees (Scalise, 2012; Shaftel, 2015) along with the potential for increased accessibility.

In summary, previous research on TEIs, which examined the degree to which TEIs measure the same construct as traditional items, indicates that TEIs are successful in presenting the same target constructs as traditional items. Moreover, the strength of TEIs rests in aspects of item design as well as psychometric properties. IRT-based studies have found that TEIs can be expected to be more difficult than traditional items, at least at some grade levels, but may have

comparable or better discrimination than traditional items. Perhaps due to the novelty of their interfaces, TEIs have been found to take longer to complete than comparable traditional items, but the information contributed to the overall test score offsets the added duration, making TEIs more efficient. Finally, although it is suggested that CBT can provide opportunities for universal tools, no studies have directly explored the effect of TEIs on test takers' use of universal tools. Further, very little empirical research has been conducted on TEIs in language assessments, and next to none has examined their use in language assessment for K–12 ELs. Thus, this study seeks to examine the TEIs used in a large-scale online language test, comparing them to equivalent traditional MCIs, to understand the effects of technological enhancements on difficulty, discrimination, efficiency, and accessibility.

3. Methods

Study Context


The current study examines TEI use in the Reading portion of ACCESS Online, a large-scale standardized ELP test used for K–12 assessment, which was developed by WIDA and the Center for Applied Linguistics. The test is annually administered to approximately 1.5 million K–12 ELs, and measures ELP in the four language domains: Listening, Reading, Speaking, and Writing. This study only concerns the Reading portion, which is administered in an online, multi-stage, adaptive format. Students are given different item sequences depending on prior performance on items, and items are grouped into three intended difficulty tiers (A, B, and C, from easiest to hardest). A distinct item bank was developed for different grade-level (G) clusters: G1, G2–3, G4–5, G6–8, and G9–12.

The test data in this study come from the 2015–2016 administration of ACCESS Online, including both operational and field test items. At that time, all operational test items were

multiple-choice items, similar to the example shown in Figure 1. However, a small number of technology-enhanced hotspot and drag-and-drop items were embedded as field test items² and presented to test takers in addition to MCIs. Hotspot items involve simple mouse click responses, similar to MCIs, but the options and distractors are embedded in images as shown in Figure 2. Drag-and-drop items involve using a mouse to drag a piece of text or media into pre-defined spaces representing response options as shown in Figure 3.

Reading Directions

Robert, Ava, and Mr. Green are reading about fish.



1 What are they reading about?

☐ Trees

☒ Fish

☐ Birds

Figure 1. *Example of a traditional multiple-choice item on ACCESS Online*

² A small set of field test items are included in each annual administration of ACCESS Online. Based on their performance, the field test items are selected to be included in subsequent administrations of ACCESS Online as operational items.

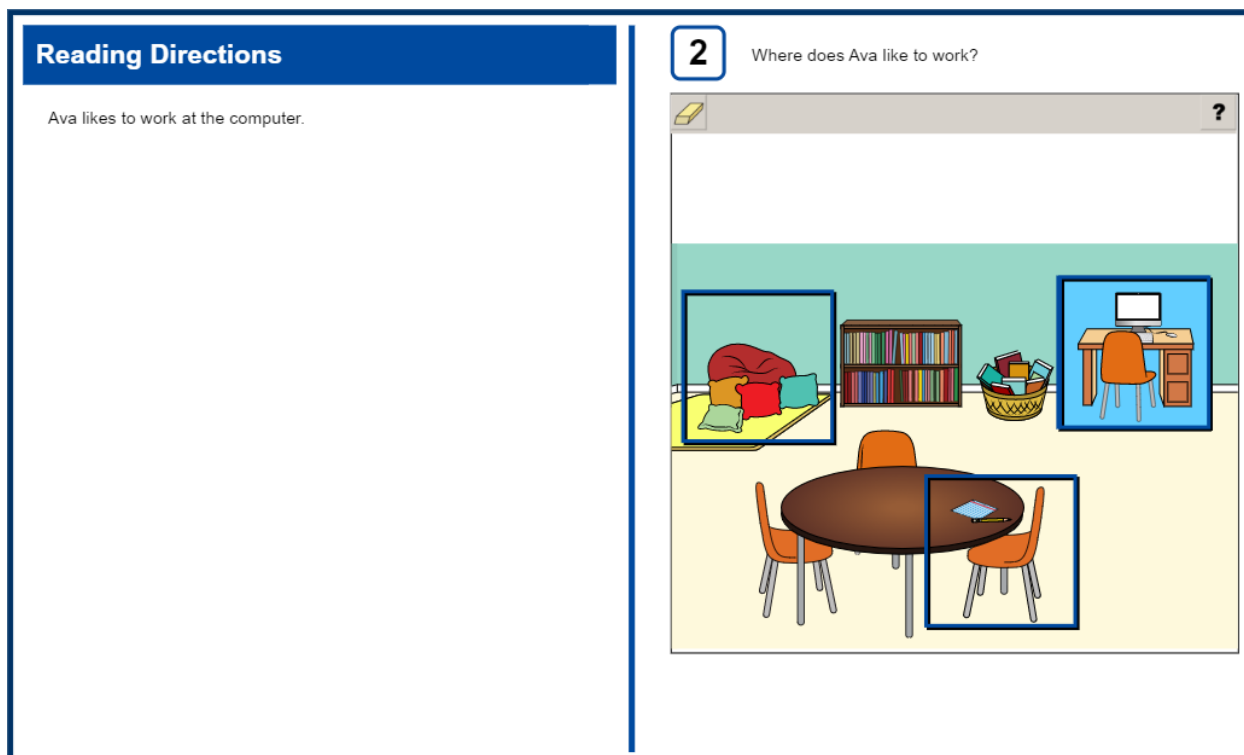


Figure 2. Example of a hotspot item on ACCESS Online

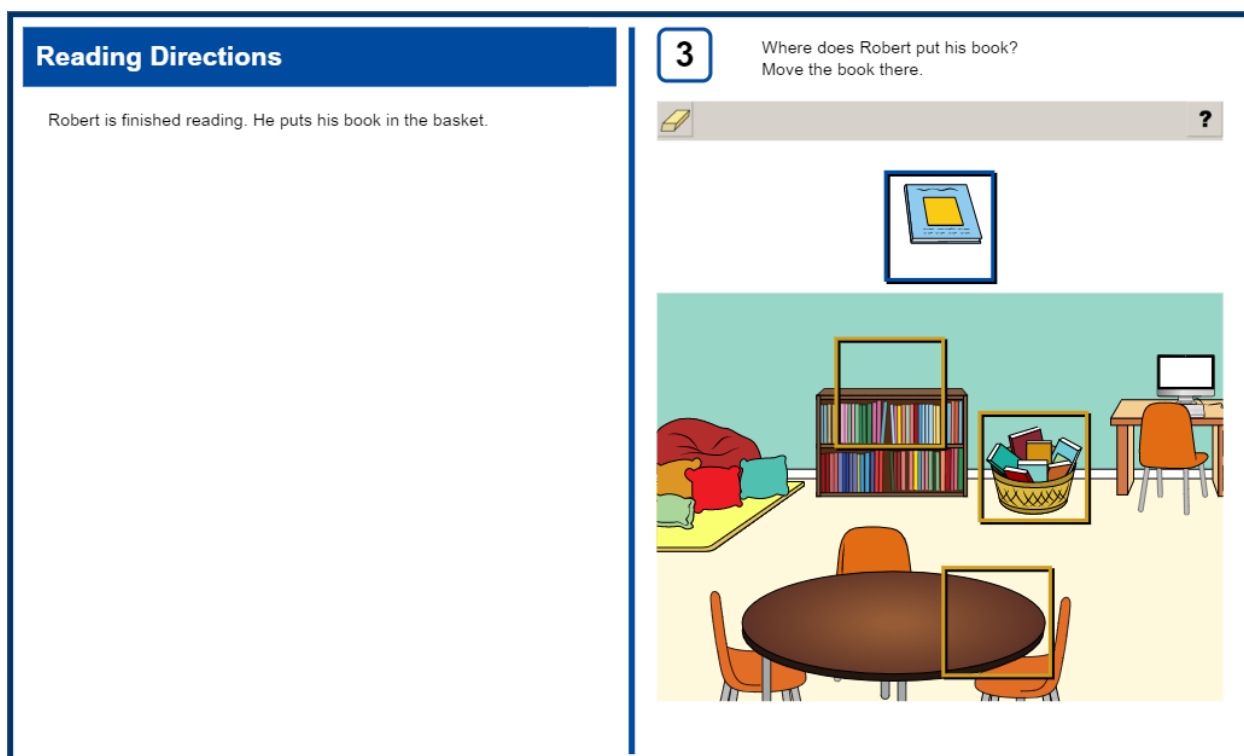


Figure 3. Example of a drag-and-drop item on ACCESS Online

By Scalise and Gifford's (2006) classification discussed above, both the hotspot and drag-and-drop items are fully constrained, similar to MCIs, but have added complexity due to the inclusion of new-media distractors (i.e., item options involving media and interactivity not available in traditional MCIs). Both TEIs and MCIs are scored dichotomously. However, the primary difference between the two item types lies in the added complexity of interpreting and selecting choices in the TEIs, which involve clicking objects (for hotspot items) or completing images via dragging (for drag-and-drop items). Table 1 gives a more complete breakdown of item dimensions for ACCESS Online MCIs and TEIs, based on Parshall et al.'s (2010) framework for classifying TEIs. Traditional MCIs, hotspot TEIs, and drag-and-drop TEIs in this study are all selected-response items. Each item type employs a graphic stimulus, but the graphics are more likely to be incorporated into the TEIs. Traditional MCIs and hotspot items have a similarly low level of interactivity, whereas the interactivity of drag-and-drop items is a little higher due to the requirement that test takers identify the potential response objects and the location of the drop area. Traditional MCIs are the least complex in terms of interface and instructions and also the most artificial in terms of action and appearance. Hotspot items are slightly more complex, with slightly higher fidelity, and drag-and-drop items are even more complex with moderate fidelity. Each item type is scored in a dichotomous objective fashion.

Table 1. *Item Dimensions for ACCESS Online MCIs and TEIs (adapted from Parshall et al., 2010)*

Item dimension	Multiple-choice items	Technology-enhanced items	
		Hotspot	Drag-and-drop
Assessment Structure	Selected Response	Selected Response	Selected Response
Response Action	Mouse click only (button)	Mouse click only (image)	Mouse click and movement
Media Inclusion	Graphics in stimulus	Graphics in stimulus and response area	Graphics in stimulus and response area
Interactivity	Low	Low	Moderate
Complexity	Low	Moderate-Low	Moderate
Fidelity	Moderate-low	Moderate	Moderate-high
Scoring	Dichotomous objective	Dichotomous objective	Dichotomous objective

In sum, the TEIs analyzed in this project are similar to traditional MCIs in that they are selected response items requiring mouse-click responses and are dichotomously scored.

However, they include different applications of media elements from the MCIs, higher fidelity to real-world problem-solving, and increased complexity. Of the two types of TEIs, the drag-and-drop items are more complex and involve more interactivity than the hotspot items.

In the ACCESS Online test interface, below the item response areas, there are universal tools that test takers can activate to enhance accessibility (see Figure 4). These features include color options (color overlay and color contrast), a highlighter tool, a line guide tool, a magnifier tool, and a help button for general and tool assistance (for definitions of the tools see Table 2). Although these tools have not been developed specifically for TEIs, test takers can activate any of these tools—and might choose to do so more frequently—when engaging with TEIs.

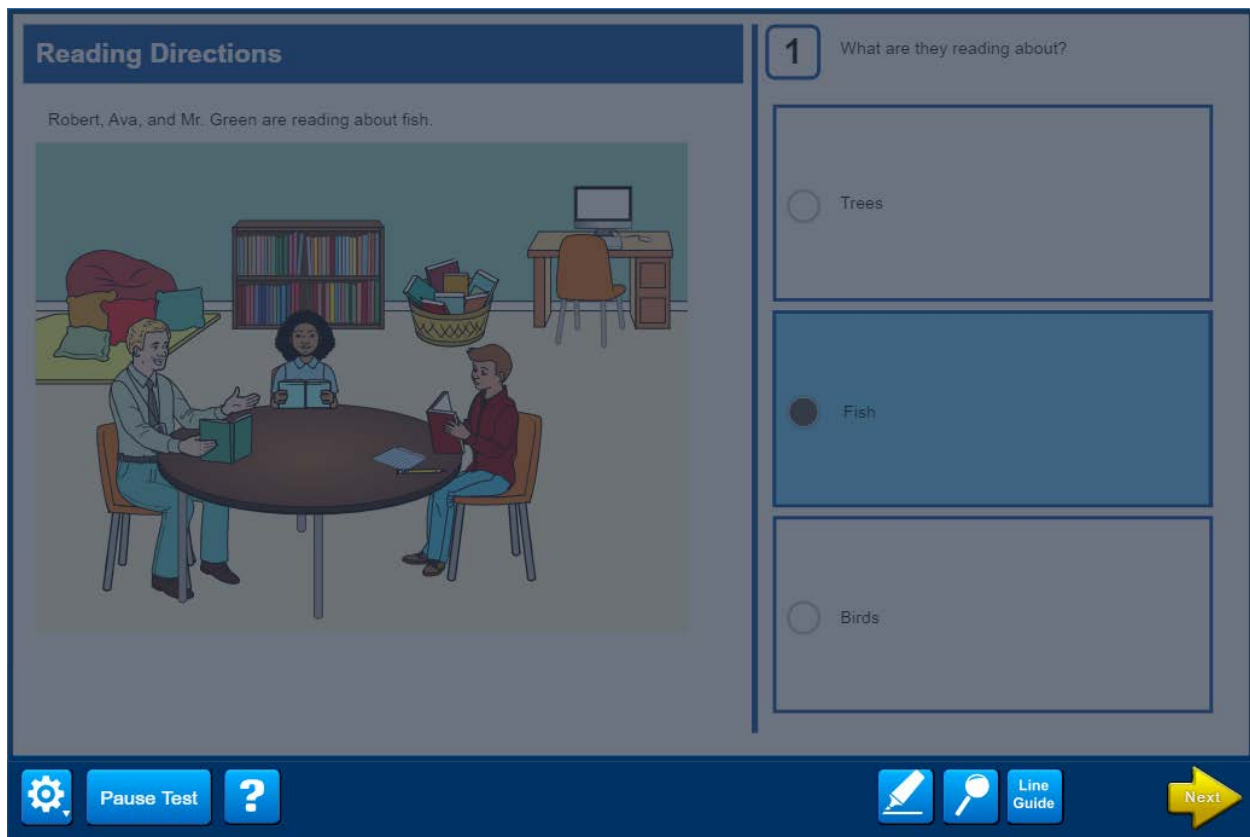


Figure 4. Universal tool bar at the bottom of the test screen.

Table 2. Description of universal tools available in the reading test

	<p>The color overlay option allows test takers to manipulate the text color and the color behind the text, graphics, and response areas. There are six options.</p>
	<p>The color contrast option allows test takers to manipulate the contrast between text and background by selecting a background color. There are six options.</p>
	<p>The highlighter tool allows test takers to mark parts of the stimulus text.</p>
	<p>The line guide tool allows test takers to drag a horizontal line across the stimulus and question text.</p>
	<p>The magnifier tool allows test takers to manipulate the graphic and text size, which can be enlarged to 1.5 or 2.0 times the default size.</p>
	<p>The help button gives test takers more information about the universal tools, with two options: a “What’s This?” feature [referred to as Help (General) in the report] that describes how to use the help tool, and an “Open Help” option [referred to as Help (Tools) in the report] that explains how to navigate the online test platform and activate the universal tools.</p>

Data

Data in this study include student test data (student demographic data, item response data) and telemetry data (indicating time spent on items and universal tool use) from the ACCESS Online Reading domain.

Student test data. A total of 1,149,027 students in grades 1 through 12 in 36 U.S. states and territories took the ACCESS Online reading test in the 2015–2016 administration. Because this study focuses on ACCESS Online, the data exclude those from the Kindergarten ACCESS for ELLs assessment, which is administered to all students in a paper format, and the ACCESS Paper assessment for grades 1–12. Approximately half of the test data were from G1–3 (see Table 3 for student distributions by grade). On the reading test, students responded to 24 to 30 operational (scored) items, all of which were MCIs, and up to six field-test items, which included both TEIs and MCIs. Each of the items was scored dichotomously. As seen in Table 3, mean scores are presented as percentage of correct answers, as the adaptive nature of the test does not require every examinee to respond to the same number of questions. For the purpose of this study, students’ response data, in the form of binary scores, were analyzed. The average score on the reading section was 52.4%. Note that students’ performance on the test is generally reported as a proficiency level (PL) of 1–6, with PL 6 being the highest. (To obtain the PL scores, students’ raw scores are converted to scale scores and then to PL scores.) The average reading proficiency level for all students was 3.57 ($SD = 1.48$).

Table 3. *Descriptive statistics for the ACCESS Online reading section*

Grade	<i>N</i>	Correct response rate (all items)		Reading proficiency level (1–6)	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
1	168,597	43.8%	15.3%	3.56	1.46
2–3	343,075	57.3%	18.5%	3.79	1.51
4–5	196,524	49.5%	16.1%	3.85	1.44
6–8	221,698	50.5%	14.1%	3.19	1.38
9–12	219,131	55.8%	15.2%	3.36	1.57
Total	1,149,025	52.4%	16.1%	3.57	1.48

The analyses in this study specifically focused on content-matched TEI and MCI pairs that were field-test items, embedded in the reading test. The content-matched item pairs shared the same topic and reading passage, and they were designed to have identical or similar prompts. The primary difference between the items in a pair was the response method. Field-test items that were not content-matched with an item of the other type were not included in the analyses. Each test taker responded to only one version of a content-matched item pair during a test administration. Operational items were analyzed along with field-test items for difficulty, discrimination, and information to provide a baseline for those measures. The number of each item type, as well as the titles of content-matched items, are presented in Table 4.

Table 4. *Number of items of each type at each grade-level cluster.*

Grade	Number of operational items for all tiers	Total number of field-test items		Content-matched field-test item pairs	Matched item pair titles
TEI	MCI				
1	72	5	19	3	Measuring with Blocks Visit to the Farmer's Market Using a Balance to Measure
2–3	72	4	16	2	Boats Long Ago New School Clubs
4–5	72	7	17	5	Buying Groceries Packing up the Library*
6–8	72	5	13	3	Measuring Angles* School Treasure Hunt Mailing Packages Surveys in Math Class
9–12	72	2	4	2	Choosing the Right College*

*Two pairs of items with this title had different content.

Telemetry data. Telemetry data is a record of individual test takers' keystrokes and mouse clicks. This information is tracked behind the scenes to better understand the activity of users in a computer-based environment. The specific telemetry data analyzed in this study includes students' use of universal tools, pauses during the test, advancement to subsequent items, as well as the time stamps of these actions. Information from clicks and screen advances can be evaluated to determine the total time spent on each test item by each student. In this study, the telemetry data for item duration and universal tool use were aggregated and analyzed.

Data Analysis

Item Response Theory (IRT) analyses. This study employs sequential explanatory design (Creswell & Plano Clark, 2018), a type of mixed-methods approach that uses both quantitative item-level measurement and qualitative analysis of item design, to address the above research questions. In preparation for examining item characteristics (research question 1), descriptive statistics of the items were collected for each grade-level cluster. To compare

performance of TEIs to traditional MCIs, we examined item parameters (difficulty and discrimination) using students' reading test scores from each grade-level cluster using a two-parameter logistic (2-PL) IRT model. IRT models create assessment scales which are robust and invariant with respect to test takers and individual items. The equation for a 2-PL model, that is, the probability of answering an item correctly based on test taker ability, is:

$$P(u = 1|\theta_i) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}$$

where for each item (i), the probability of a correct response is modeled by the variable parameter of ability (θ) and the constant parameters of item difficulty (b) and item discrimination (a). The *item characteristic curves* (ICC) from the IRT model (see Figure 5, in grey) represent the probability of answering an item correctly given an individual's ability, the item difficulty, and the item discrimination.

In IRT, the *item information function* accounts for the expectation that reliability varies relative to person ability; that is, items are not equally discriminating across all test taker ability levels. This relative measure of reliability is known as Item Information (I), and it is calculated thusly:

$$I_i(\theta, b_i, a_i) = a_i^2 P_i(\theta, b_i) Q(\theta, b_i)$$

where for a given item (i), the information is modeled across person ability (θ), with constant parameters for item difficulty (b) and item discrimination (a), given the probability of a correct answer (P) and the probability of an incorrect answer (Q).

Item information values are useful for visualizing the contribution of an item to the overall effectiveness of a test at different ability levels (see Figure 5, in black). Typically, curves with higher peaks indicate items are more informative (i.e., provide reliable scores across ability levels). The position of the peak of this curve along a range of abilities (x-axis) indicates the

level of ability for which the item is most informative (the parameter b). Taken together, both visual and numeric results from IRT models provide rich information about the performance of items on an assessment. In the example in Figure 5, the item information is shown across ability levels, with peak information around $-.25$. For the item represented by the figure, this means the most reliable measurement of the latent construct occurs for test takers with an ability level of $-.25$, because at that ability level, the item characteristic curve shows a 50% of a test taker getting the item correct.

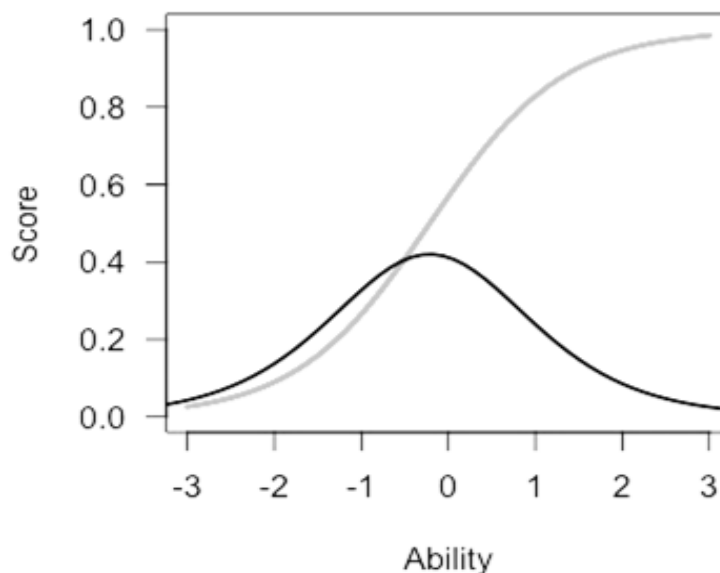


Figure 5. *Example of an item characteristic curve (grey) and an item information curve (black)*

Item efficiency. To analyze the efficiency of TEIs against MCIs for research question 2, the time test takers spent on each item (i.e., item duration) was compared using telemetry data. First, descriptive statistics of average item duration for each item type in each grade-level cluster are presented. To compare the item duration of TEIs and MCIs, Wilcoxon's W was measured due to the non-normality of the data; effect sizes (r) were obtained as well. In addition, item efficiency was calculated for each item, based on the ratio of item information to item duration per minute. This metric is useful for interpreting differences in average duration, as an item

having a relatively higher item duration might not be undesirable if the item efficiency is also high.

Accessibility and universal tools. The accessibility of TEIs as compared to MCIs, the focus of research question 3, was approached from the perspective of universal design. We explored whether TEIs led to more activation of the universal tools embedded in the test platform. Unexpected overuse of the tools on a particular item might indicate an accessibility issue with that item. The null hypothesis would be that observed tool use is evenly distributed between content matched TEIs and MCIs. Frequency of universal tool use for TEIs and MCIs were organized into contingency tables, which included the baseline assumption that tool use would be equivalent in the two item types. The data are observations of occurrences rather than ratios, and there are no expectations about how frequently tools will be used across item types or how much tool use is appropriate. Thus, these tables were analyzed using chi-square goodness-of-fit tests for each item within each grade-level cluster to show which type of item elicits more use of universal tools.

Item design analysis. For each research question, the researchers performed a qualitative examination of item content to understand why certain TEIs performed differently than the paired traditional MCI based on the above analyses. Language assessment research has long relied on thorough content analysis (Bachman, Davidson, & Milanovic, 1996; Bachman, Kunnan, Vanniarajan, & Lynch, 1988) to understand the influence of test content on performance. This approach has not been strictly applied to the construct of technological enhancements, as researchers often do not have access to content-matched pairs of TEIs and MCIs.

As ACCESS Online is designed with a commitment to universal design principles (CAST, 2011; Hall, Meyer, & Rose, 2012; National Center on Universal Design for Learning, 2012), specific attention is given to aspects of the TEI interfaces that relate to the universal design principles of *means of representation*, *action and expression*, and *engagement*. Providing means of representation refers to offering multiple channels of information, using multimedia, clarifying language and symbols, and guiding processing. Providing means of action and expression refers to offering assistive technology (e.g., universal tools), varying response and communication methods, and guiding strategy use. Providing means of engagement refers to maximizing interest, minimizing distractions, making objectives salient, and promoting motivation.

The content analysis began with the researchers' identifying the elements or qualities of each TEI that was distinct from its MCI counterpart. These distinctions were then categorized as changes in response actions; changes in content; and gains or losses in means of representation and engagement, based on principles of Universal Design for Learning (National Center on Universal Design for Learning, 2012). From the TEIs that met the criteria of this last category, we selected two to examine closely in light of the results to each research question. We then provide detailed explanations of which distinctions between paired TEIs and MCIs may lead to differences in test taker performance and affect the authenticity of the items (Russell, 2016).

4. Findings

Findings on research question 1: Item Characteristics

To examine the item characteristics of TEIs as compared with MCIs (i.e., item difficulty and discrimination), 2-PL IRT analyses were conducted on content-matched pairs of TEIs and MCIs. The performance of each of these items was compared to the average performance of

items on the entire test. The discrimination (a) value reflects the relative reliability of the item, and the difficulty (b) value reflects the test taker ability level at which the item had a 50% chance of being responded to correctly. The discrimination value can theoretically range from 0 to infinity, as it is the slope of the item curves, with higher values indicating items that contribute more information to the overall test (i.e., are more reliable and discriminating). In this analysis, a values ranged from 0.1 to 3.58. The b -parameter represents item difficulty, with lower measures (including negatives) indicating easier items and higher measures indicating more difficult items. Difficulty relates to item information because higher b values indicate the item provided more information with higher-level examinees' responses to the item. These values can also theoretically range from negative to positive infinity, but in this analysis are typically within a -3 to +3 range.

Table 5 presents item discrimination and difficulty parameter values for each of the content-matched TEI and MCI pairs, as well as averages of all TEIs, all MCIs, and all items on the whole test. For example, in G1 Pair 1 ("Measuring with blocks"), the TEI hotspot version of the item had a higher item difficulty value of 1.189 compared to the MCI which had a difficulty of .600, suggesting that more students responded correctly to the MCI than to the TEI. On the same item, item discrimination for the TEI (.551) was higher than that for the MCI (.431), indicating that the TEI was more discriminating than the MCI.

Table 5. *Item Characteristics*

Grade	Item Set (item topic)	Type*	Discrimination (<i>a</i>)	Difficulty (<i>b</i>)
1	Pair 1	HS	0.551	1.189
	(“Measuring with blocks”)	MC	0.431	0.600
	Pair 2	HS	0.710	-0.532
	(“Farmer’s Market”)	MC	0.746	-0.974
	Pair 3	DD	0.372	0.188
	(“Using a balance”)	MC	0.675	-0.712
	Average of TEIs (n = 3)	TEI	0.539	0.428
	Average of MCIs (n = 3)	MC	0.590	-0.224
	Overall average (n = 78)**		0.925	-1.093
2–3	Pair 1	HS	1.096	-0.126
	(“Boats”)	MC	1.210	-0.062
	Pair 2	HS	0.769	-1.453
	(“School Clubs”)	MC	0.997	-1.343
	Average of TEIs (n = 2)	TEI	1.088	-0.160
	Average of MCIs (n = 2)	MC	1.205	-0.095
	Overall average (n = 76)**		1.211	-0.483
4–5	Pair 1	DD	2.128	-0.202
	(“Groceries”)	MC	2.080	-0.361
	Pair 2	DD	0.414	0.604
	(“Library 1”)	MC	0.266	-0.226
	Pair 3	DD	0.620	-0.420
	(“Library 2”)	MC	0.523	-0.689
	Pair 4	HS	1.175	-0.085
	(“Measuring Angles 1”)	MC	1.079	-0.464
	Pair 5	HS	0.698	3.625
	(“Measuring Angles 2”)	MC	0.744	3.345
	Average of TEIs (n = 5)	TEI	0.890	1.434
	Average of MCIs (n = 5)	MC	0.846	1.067
	Overall average (n = 82)**		1.041	1.094
6–8	Pair 1	DD	0.571	0.854
	(“Treasure Hunt”)	MC	0.215	-0.255
	Pair 2	HS	0.914	1.808
	(“Mailing Packages”)	MC	1.030	1.178
	Pair 3	HS	1.132	2.274
	(“Surveys”)	MC	1.075	1.303
	Average of TEIs (n = 3)	TEI	0.900	2.121
	Average of MCIs (n = 3)	MC	0.860	0.909
	Overall average (n = 78)**		1.157	1.294
9–12	Pair 1	HS	1.038	2.914
	(“Choosing college 1”)	MC	1.052	1.230
	Pair 2	DD	0.969	2.615
	(“Choosing college 2”)	MC	1.035	2.451
	Average of TEIs (n = 2)	TEI	1.003	2.727
	Average of MCIs (n = 2)	MC	1.044	1.758
	Overall average (n = 76)**		1.166	1.519

*Note: HS = hotspot item; MC = multiple-choice item; DD = drag-and-drop item; TEI = technology-enhanced item (i.e., hotspot items and drag-and-drop items)

**“Overall average” refers to the average scores of content-matched TEIs and MCIs along with the 72 operational MCIs on the test for each grade-level cluster.

Overall, TEIs were more difficult than MCIs as seen in Figure 6, which displays the relative difficulty levels for TEIs, MCIs, and all test items for each grade-level cluster. Only at G2–3 were the TEIs, on average, less difficult than the content-matched MCIs, likely due to one pair of items (“Boats”) in which the MCI had a higher difficulty value than the content-matched TEI. When compared to the overall test difficulty, the difficulty of TEIs was most similar in G2–3 and G4–5, but farther from that of the overall test in other grade-level clusters. Although TEIs were often more difficult than their matched MCIs, the most difficult TEI was paired with the most difficult MCI, indicating that the content may have affected the difficulty of the items. Within TEIs, hotspot items were generally more difficult than drag-and-drop items, and the hotspot items had a larger range of difficulty levels. This is in line with initial cognitive labs conducted on ACCESS Online TEIs, in which hotspot items were found to pose more of a challenge to learners than drag-and-drop items (CAL, 2015).

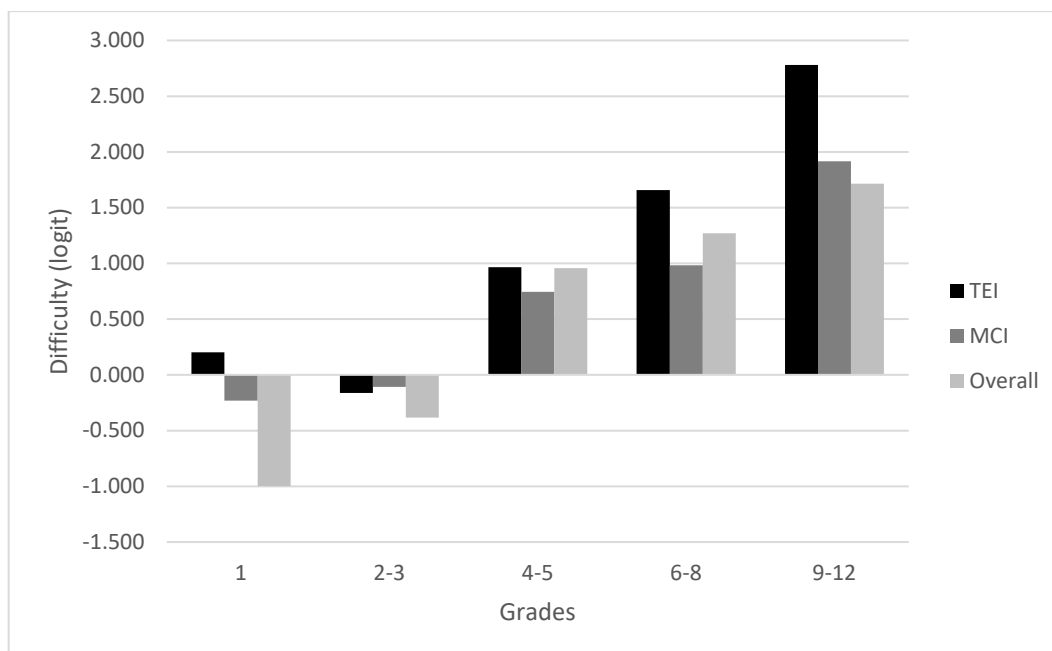


Figure 6. *Average item difficulty within item types for each grade-level cluster on a logit scale.*

The discrimination values of TEIs varied across grade-level clusters. TEIs provided comparable but lower levels of discriminative power for G1 (TEI, .539; MCI, .590), G2–3 (TEI, 1.088; MCI, 1.205), and G9–12 (TEI, 1.003; MCI, 1.044). However, TEIs provided slightly higher discriminative power than MCIs for G4–5 (TEI, .890; MCI, .846) and G6–8 (TEI, .900; MCI, .860). It is important to note that discrimination values for TEIs and MCIs followed similar trends across items. In other words, if discrimination for an MCI is relatively high, as compared to other MCIs, its content-matched TEI is likely to have higher discrimination than other TEIs as well. This pattern suggests that content may influence item discrimination more than item type.

Item information curves show that any given item is unlikely to be equally informative at all levels of test taker ability, and we can interpret the discriminative power of an item as it relates to those ability levels. TEIs and MCIs, on average, provided similar levels of information to the overall test score, though TEIs and MCIs were often most informative at distinct ability levels. Visual representations of average item information are presented by grade-level cluster in Figures 7-11. In these graphs, item information curves are plotted along person-ability on the x-axis, with information (i.e., measurement accuracy) increasing on the y-axis. The higher the curve, and the more space beneath the curve, the more discriminating the item is and the more information that item provides to the overall score. Separate curves are presented for TEIs, for MCIs, and for the overall test. These curves are scaled for the number of items included in the measurement on the curve so that visual comparison is meaningful. In addition, the vertical lines divide the graphs into six proficiency levels (ranging between 1 and 6) to present the informativeness of items per PL.

Findings show that TEIs were more informative than MCIs in only select grade-level clusters: G4–5 and G6–8 (see Figures 7-11). In addition, at G4–5, TEIs were most informative at

PL 2 along with the overall test (see Figure 9). At G6–8, TEIs were most informative at PL 4, whereas the MCIs and the overall test were most informative around the cutoff between PL 2 and PL 3 (see Figure 10). At G1, although TEIs were generally less informative than MCIs, TEIs were slightly more informative for students at PL 6 (see Figure 7). Likewise, at G9–12, TEIs were most informative at PL 5, whereas the MCIs and overall test were most informative around the cutoff between PL 2 and PL 3 (see Figure 11).

The above findings suggest that the informativeness of TEIs and MCIs varied not only by grade, but also by test taker proficiency level. For G1 and G2–3, it is clear that TEIs were less informative than MCIs. Additionally, at these grade levels, TEIs provided the most information at the same proficiency level as MCIs; that is, information curves for both sets of items reached a peak at similar proficiency levels. However, at G6–8 and G9–12, although TEIs were slightly less informative than MCIs, they provided better information about test takers at higher ability levels, indicating a unique contribution to the overall test score.

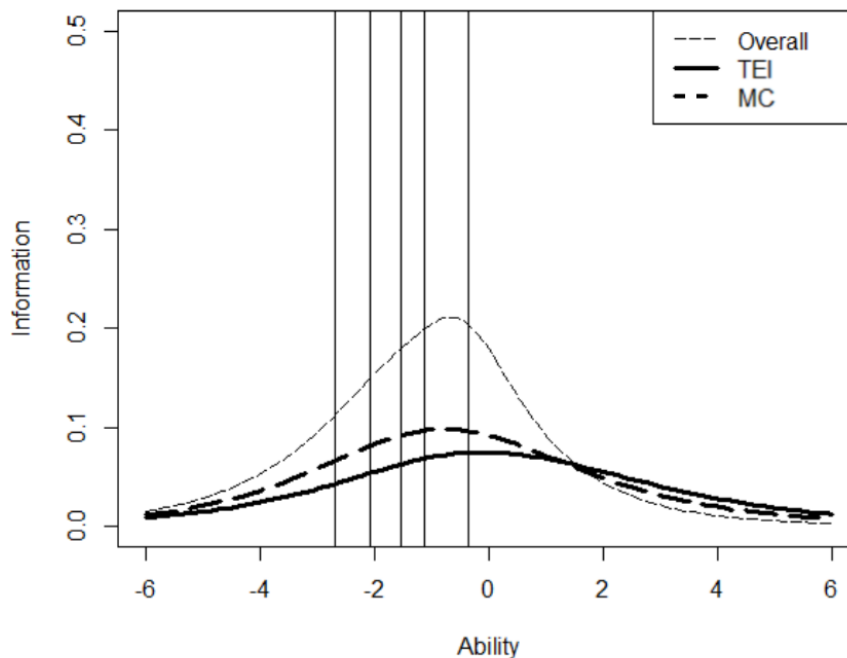


Figure 7. *Item information curves for grade 1 items.*

Note: Vertical lines represent proficiency level cut scores, and the spaces around the cut scores represent the six proficiency levels. For meaningful visual comparison, the area under each curve is normalized for the number of items in each category: 3 TEIs, 3 MCIs, 78 items overall.

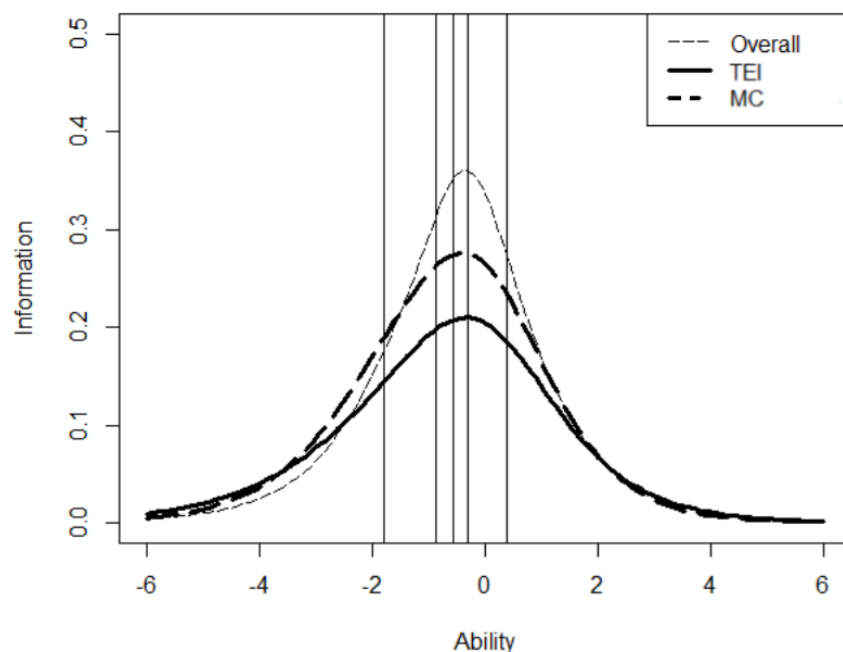


Figure 8. *Item information curves for grade 2–3 items.*

Note: Vertical lines represent proficiency level cut scores, and the spaces around the cut scores represent the six proficiency levels. For meaningful visual comparison, the area under each curve is normalized for the number of items in each category: 2 TEIs, 2 MCIs, 76 items overall.

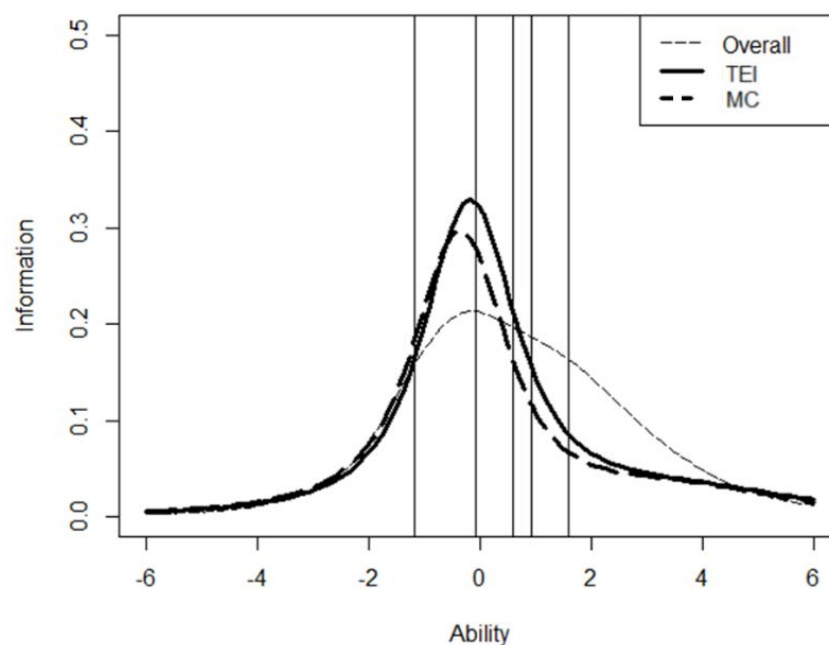


Figure 9. *Item information curves for grade 4–5 items.*

Note: Vertical lines represent proficiency level cut scores, and the spaces around the cut scores represent the six proficiency levels. For meaningful visual comparison, the area under each curve is normalized for the number of items in each category: 5 TEIs, 5 MCIs, 82 items overall.

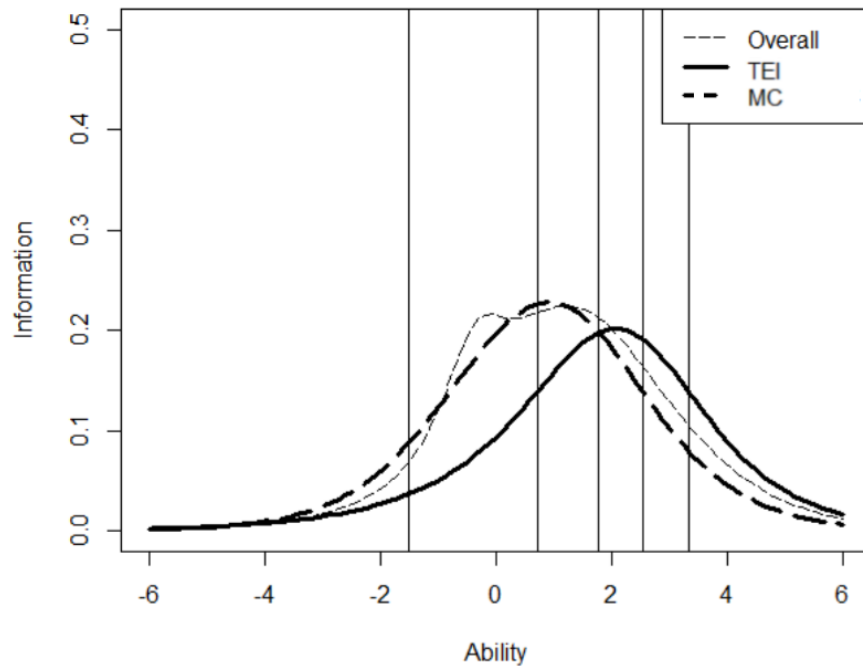


Figure 10. *Item information curves for grade 6–8 items.*

Note: Vertical lines represent proficiency level cut scores, and the spaces around the cut scores represent the six proficiency levels. For meaningful visual comparison, the area under each curve is normalized for the number of items in each category: 3 TEIs, 3 MCIs, 78 items overall.

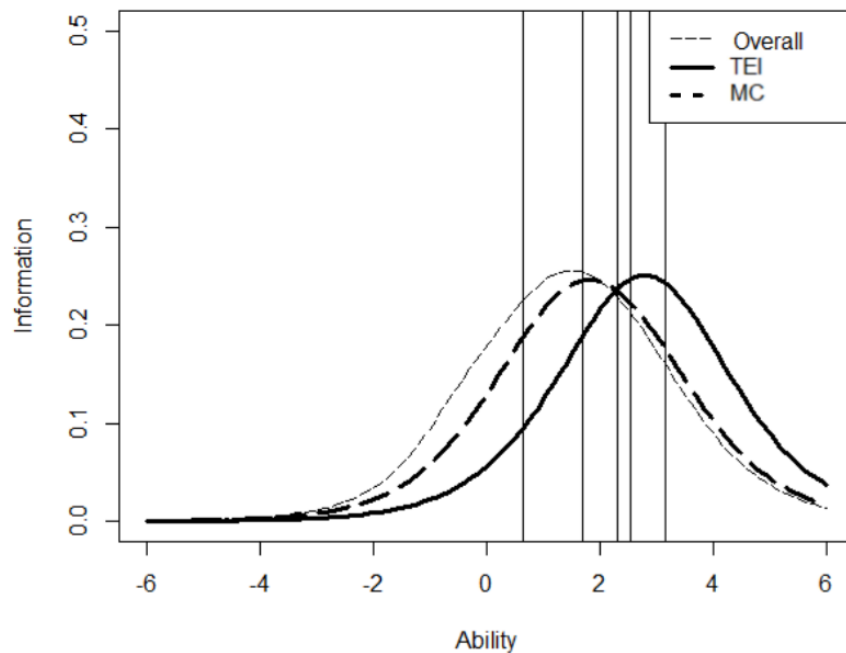


Figure 11. *Item information curves for grade 9–12 items.*

Note: Vertical lines represent proficiency level cut scores, and the spaces around the cut scores represent the six proficiency levels. For meaningful visual comparison, the area under each curve is normalized for the number of items in each category: 2 TEIs, 2 MCIs, 78 items overall.

Quantitative findings indicated that TEIs were more difficult than MCIs, and TEIs offered more information than MCIs about test takers at high proficiency levels. Additional in-depth qualitative analyses were conducted to identify design features which may have contributed to the increased difficulty. Most TEIs involved removal of some content or information, even in a small way, a change which might reduce the means of representation of information in an item but which might also make the item more discriminating among test taker ability levels.

To illustrate, the G4–5 item pair “Library 2” involved answering a question about shelving books on a bookshelf. In both the MCI and TEI, test takers had to identify how many books would fit on a bookshelf based on information in the reading passage. In this case, the TEI version was more difficult but also more discriminating and informative. The MCI version had answer choices that were arithmetic expressions, one of which correctly described how many books fit on a bookshelf. The response options included representations of both multiplication and division. In the drag-and-drop TEI version, each answer choice was instead a single numeral which could be dragged into a box to complete a partially written-out arithmetic expression. In addition to the change in interface, the exact numeric options were changed and only multiplication was represented. This change, in a sense, reduces the means of representation in the TEI version, as the different arithmetic operators might be a meaningful source of information that test takers use too solve the problem in the MCI.

The only pair in which the TEI was less difficult than the traditional MCI, the G2–3 item pair “Boats,” involved answering a question about the history of boats. It was, however, less discriminating than the traditional MCI version. The answer choices in the TEI were pictures of boats, and the correct choice was described in the passage. In both the MCI and TEI, the

response options were presented as pictures. In the MCI version, the answer option pictures were presented next to clickable buttons, whereas in the TEI version, the pictures were themselves clickable hotspot options. Regarding differences in item content presentation, the TEI version had the options presented as a table, which had a title not present in the MCI. Overall, the TEI version gives test takers more textual information as well as multiple means of interpreting the information in the passage. The TEI increases the means of representation, in terms of universal design (CAST, 2011; National Center for Universal Design for Learning, 2012), which might contribute to the decreased difficulty.

Interpretations of TEIs should be made cautiously for lower grades and proficiency levels, where construct-irrelevant factors, such as computer literacy and general test familiarity can significantly influence performance. It is important to consider that item content also exerted a strong influence on trends in item difficulty and information as item type. Most content-matched item pairs were more similar to each other in performance than to other item pairs across the measurements (see Table 5). For example, at G2–3, Pair 1 (“Boats”) had higher difficulty values than that of the overall test, and the two items in the pair were much closer to each other in terms of discrimination and information than they were to the same-type items in Pair 2 (“School Clubs”), which were relatively easier and less informative. This finding indicates that the content distinction between “Boats” and “School Clubs” had a larger impact on item characteristics than the response format. Still, the results indicate that care should be taken when developing innovative, technology-enhanced item types for younger and less proficient populations.

Findings on research question 2: Efficiency of TEIs vs. MCIs

To understand efficiency of TEI reading items, the focus of research question 2, telemetry data on time spent to respond to each item were analyzed. These data were used to compare the average item durations of TEIs as compared to MCIs. Item duration was also used to calculate the efficiency of each item—that is, item-level information divided by average duration.

Item duration of TEIs and MCIs. The first aspect of research question 2 pertains to the difference in item duration (i.e., the time a test taker spends responding to an item) between TEIs and MCIs. Table 6 presents descriptive statistics for content-matched TEIs and MCIs at each grade-level cluster in terms of item duration (in seconds). Figure 12 plots average item duration for TEIs and MCIs at each grade-level cluster. Item type is indicated, as well as the number of responses per item.

At every grade-level cluster, TEIs on average required more time to answer than traditional MCIs. This held true for most content-matched pairs of items, with the exceptions of the G1 Pair 1 items, the G4–5 Pair 5 items, and the G6–8 Pair 2 and Pair 3 items. For these item pairs, MCIs took on average a couple of seconds longer to answer than TEIs. However, for all other item pairs, the TEIs had longer item durations. In one extreme case, G9–12 Pair 2, the average duration for the TEI was more than a minute longer than that of the MCI. In addition, the difference between TEIs and MCIs in item duration increased with each grade-level cluster, except for G6–8, which had the smallest average difference in duration between item types.

Table 6. *Descriptive statistics for item duration on TEIs vs. MCIs*

Grade	Item Set	Item Type*	N	Mean (sec.)	SD (sec.)
1	Pair 1	HS	25710	36.840	34.237
		MC	25712	37.983	33.957
	Pair 2	HS	15849	33.871	37.385
		MC	16127	32.311	34.323
	Pair 3	DD	18136	41.848	39.955
		MC	18574	32.595	27.304
	Total	TEI	59695	37.573	36.890
		MC	60413	34.812	32.162
2–3	Pair 1	HS	69985	42.638	31.525
		MC	70426	38.300	27.336
	Pair 2	HS	1831	70.910	101.307
		MC	1843	36.272	32.351
	Total	TEI	71816	43.359	35.072
		MC	72269	38.249	27.470
4–5	Pair 1	DD	2559	50.923	49.824
		MC	2516	30.323	24.289
	Pair 2	DD	9811	44.051	33.321
		MC	9928	35.235	24.407
	Pair 3	DD	9958	66.635	53.608
		MC	9842	48.450	41.190
	Pair 4	HS	45533	42.933	28.604
		MC	44874	36.335	26.039
	Pair 5	HS	45638	50.380	37.447
		MC	44962	54.966	45.275
	Total	TEI	113499	54.911	36.002
		MC	112122	40.518	36.171
6–8	Pair 1	DD	19760	74.382	69.589
		MC	20085	56.810	52.743
	Pair 2	HS	43814	64.634	42.664
		MC	43462	67.321	47.250
	Pair 3	HS	25449	76.526	56.701
		MC	25712	77.471	61.130
	Total	TEI	89023	70.197	53.757
		MC	89259	67.880	52.815
9–12	Pair 1	HS	31788	83.911	61.103
		MC	32134	59.621	44.556
	Pair 2	DD	32053	168.528	158.479
		MC	32233	95.820	70.315
	Total	TEI	63841	126.395	120.287
		MC	64367	77.748	58.881

*Note: HS = hotspot item; MC = multiple-choice item; DD = drag-and-drop item; TEI = technology-enhanced item (i.e., hotspot items and drag-and-drop items)

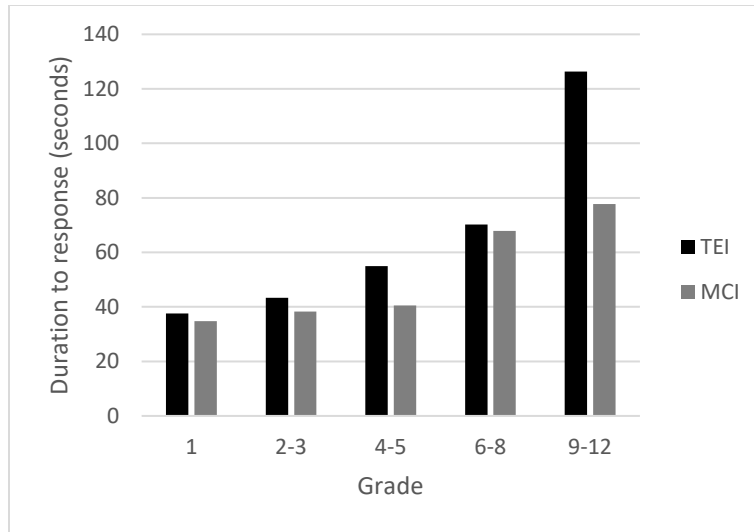


Figure 12. *Average item duration for each grade-level cluster in seconds*

To determine whether the difference in item duration between TEIs and MCIs was meaningful, a series of Wilcoxon’s *W* unpaired tests were carried out due to the large standard deviations of the duration measurements. These large standard deviations indicate a strong skew, suggesting the need for non-parametric tests. The Wilcoxon test functions similarly to a standard *t*-test, comparing a numeric variable measured for two categorical groups, but this test can be applied to non-normally-distributed data like these item durations. In Table 7, each comparison is listed along with the test statistic (*W*), *p*-value, and effect size (*r*). Due to the very large sample size, the majority of the item pair comparisons were significant; therefore, it was necessary to interpret the effect sizes to establish the meaningfulness of the item duration differences. For most item pair comparisons, the effect size of the difference between the two item types was small ($.1 < r < .3$) or negligible ($r < .1$). The exception was the G4–5 Pair 1 items, which had an effect size of .402. In general, effects were stronger between drag-and-drop items and traditional MCIs than between hotspot items and traditional MCIs. Drag-and-drop items were part of G1 Pair 3, G4–5 Pairs 1–3, G6–8 Pair 1, and G9–12 Pair 2 items. Average item duration differences

between all TEIs and paired traditional MCIs were statistically negligible in every grade-level cluster, except for at G9–12, where there was a small effect of the average difference in duration.

Table 7. *Wilcoxon's W (unpaired) tests of item duration differences between matched TEIs and MCIs.*

Grade	Item	Comparison*	W	p	r
1	Pair 1	HS – MC	334950386	0.007	-0.012
	Pair 2	HS – MC	124374174	< 0.001	0.023
	Pair 3	DD – MC	135879902	< 0.001	0.167
	Group	TEI – MC	1700840278	< 0.001	0.049
2–3	Pair 1	HS – MC	2203646544	< 0.001	0.092
	Pair 2	HS – MC	1158592	< 0.001	0.271
	Group	TEI – MC	2306009946	< 0.001	0.096
4–5	Pair 1	DD – MC	1723992	< 0.001	0.402
	Pair 2	DD – MC	38316913	< 0.001	0.185
	Pair 3	DD – MC	35732290	< 0.001	0.234
	Pair 4	HS – MC	824927901	< 0.001	0.128
	Pair 5	HS – MC	1113388343	< 0.001	-0.074
	Group	TEI – MC	5821712260	< 0.001	0.074
6–8	Pair 1	DD – MC	156679826	< 0.001	0.182
	Pair 2	HS – MC	990516594	< 0.001	-0.035
	Pair 3	HS – MC	321279825	< 0.001	-0.015
	Group	TEI – MC	3823978097	< 0.001	0.032
9–12	Pair 1	HS – MC	344801216	< 0.001	0.152
	Pair 2	DD – MC	304043428	< 0.001	0.151
	Group	TEI – MC	1369853836	< 0.001	0.107

Note: Effect sizes (r) range from -1 to 1, with effect sizes between $\pm .1$ and $\pm .3$ considered small, $\pm .3$ and $\pm .5$ medium, and beyond $\pm .5$ strong.

*Note: HS = hotspot item; MC = multiple-choice item; DD = drag-and-drop item; TEI = technology-enhanced item (i.e., hotspot items and drag-and-drop items)

Efficiency of TEIs and MCIs. Some items that take more time to complete might also provide more information, so average item durations were divided by item information to produce a ratio quantifying item efficiency. Following the convention in previous research (Crabtree, 2016; Jodoin, 2003; Wan & Henly, 2012), item efficiency was measured as information per minute. In essence, items which have higher informativeness and lower

durations will have the highest measurements of efficiency. Item efficiency measurements are presented in Table 8 and in Figure 13. Efficiency ratings ranged from .227 to 4.115 information logits per minute. The average efficiency across all items was 1.09 logits of information per minute.

Table 8. Item efficiency for each item at each grade-level cluster.

Grade	Item Set	Type*	Mean Item Duration (min.)	Item Efficiency (Information per minute)
1	Pair 1	HS	0.61	0.897
		MC	0.63	0.680
	Pair 2	HS	0.56	1.259
		MC	0.54	1.385
	Pair 3	DD	0.70	0.533
		MC	0.54	1.242
	Group	TEI	0.63	0.860
		MC	0.58	1.017
2–3	Pair 1	HS	0.71	1.543
		MC	0.64	1.895
	Pair 2	HS	1.18	0.651
		MC	0.60	1.650
	Group	TEI	0.72	1.505
		MC	0.64	1.889
4–5	Pair 1	DD	0.85	2.507
		MC	0.51	4.115
	Pair 2	DD	0.73	0.564
		MC	0.59	0.453
	Pair 3	DD	1.11	0.558
		MC	0.81	0.647
	Pair 4	HS	0.72	1.642
		MC	0.61	1.781
	Pair 5	HS	0.84	0.831
		MC	0.92	0.813
	Group	TEI	0.92	0.973
		MC	0.68	1.253
6–8	Pair 1	DD	1.24	0.461
		MC	0.95	0.227
	Pair 2	HS	1.08	0.848
		MC	1.12	0.918
	Pair 3	HS	1.28	0.888
		MC	1.29	0.833
	Group	TEI	1.17	0.769
		MC	1.13	0.760
9–12	Pair 1	HS	1.40	0.742
		MC	0.99	1.059
	Pair 2	DD	2.81	0.345
		MC	1.60	0.648
	Group	TEI	2.11	0.476
		MC	1.30	0.805

*Note: HS = hotspot item; MC = multiple-choice item; DD = drag-and-drop item; TEI = technology-enhanced item (i.e., hotspot items and drag-and-drop items)

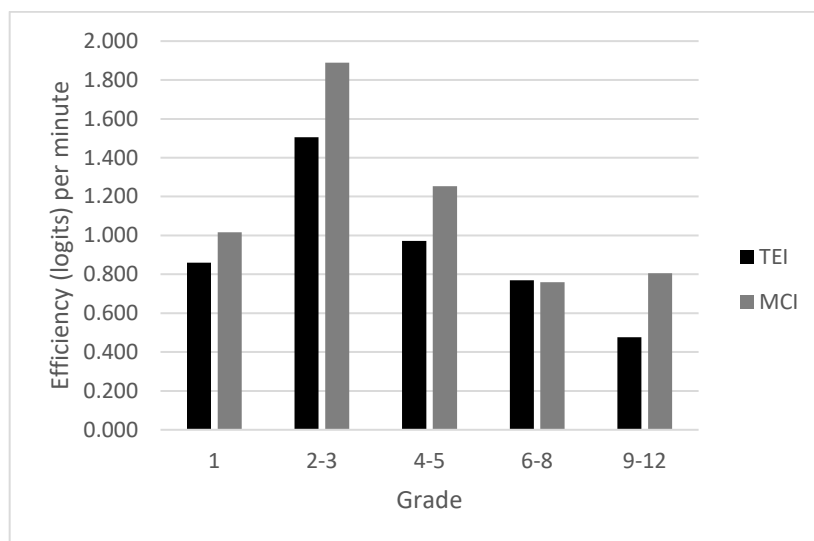


Figure 13. *Efficiency of items in each grade-level cluster in terms of information per minute.*

Most TEIs were less efficient than their traditional MCI counterparts, except for at G6–8. For this grade-level cluster, Pair 1 and Pair 3 were more efficient than traditional MCIs, but the differences were small. Between the two TEI types of hotspot and drag-and-drop, there was no distinct trend indicating which item type was more efficient.

Overall, TEIs typically took longer to complete than MCIs and were often less efficient, meaning the difference in duration was not often balanced by additional information. Design features which might have contributed to the increased duration and decreased efficiency of TEIs were again examined through qualitative analysis of the items. In addition to removal of some content as previously discussed, TEI versions could include different response instructions and content layouts. In some cases, the content arrangement and response format were more distracting and unclear than in the traditional MCI. In others, the TEI had removed repetitive, redundant, or possibly distracting content.

In the previously mentioned G4–5 item “Library 2,” the TEI version removed some helpful information; the inclusion of multiple mathematical operations in the MCI version

provided multiple sources of information that test takers could use to answer the question. The TEI version was more difficult, yet also more discriminating; the change in representation of information might have been what caused the increase in time duration. In contrast, in the item pair “Surveys” at G6–8, the TEI version was again more difficult and more discriminating, but in this case did not take any longer, on average, for a test taker to complete the TEI than to complete the MCI. It was also a math-focused item, and it asked test takers to interpret pie charts and read a short passage about a survey to answer a question about the ratio of survey responses. In the MCI version of this item, a reading passage and series of pie charts were presented as part of the prompt, and answer choices were numeric fractions. In the TEI version, the pie charts from the prompt area were moved to the answer area and presented as hotspot items. Although this removed some information that could be useful to test takers, it ultimately reduced redundant and potentially distracting information, making a clearer and more engaging item. Thus, this item’s TEI version was not only more difficult and more discriminating, it also had a lower average item duration than the MCI. As it incorporated the real-world task of examining the pie charts directly into the answer selection process, the TEI version arguably increased the fidelity of the item, and the removal of redundant information likely added clarity, and thus increased engagement. In sum, TEIs tend to increase item duration, but in the case that the TEI format allows for the removal of redundant information and provides a clear response process, item duration is not at risk of changing drastically.

Findings on research question 3: Accessibility of TEIs vs. MCIs

To understand the accessibility of TEIs, the focus of research question 3, telemetry data on use of universal tools were analyzed. Accessibility in this study was measured through test takers’ rates of activation of universal tools while responding to test items, as use of tools might

signal issues in approaching an item. Table 9 presents the total number of activations of each universal tool for each content-matched pair of TEIs and MCIs. Differences in activation rate between TEIs and MCIs were compared using a chi-square goodness-of-fit test for each pair of items. The default assumption was that TEIs and MCIs elicit universal tool activation equally, so a goodness-of-fit test was employed to determine whether either version of any item pair elicited significantly more tool activation. Numbers in bold in Table 9 represent tool activation that was significantly higher for one item type than the other within a pair of content-matched items, given that there were at least five activations of the tool for an item (an assumption of the goodness-of-fit test).

The highlighter, magnifier, and line guide tools were more frequently activated than the help and color tools. Of these, the highlighter was most-used across all grade-level clusters, with the exception of G1, where the magnifier was the most activated tool. When there was a significant difference in activation of universal tools between the two types of items in a pair, the TEI version elicited more universal tool use. TEIs elicited a significantly higher proportion of highlighter activations than the MCIs, except for the items in G1 Pair 1 and Pair 2 and in G4–5 Pair 3. The magnifier was activated significantly more often in the TEI version of the items in G4–5 Pair 3 and Pair 4 and in G9–12 Pair 2. The line guide was activated significantly more often in the TEI version of the G2–3 Pair 2 items, the G4–5 Pair 4 items, and the G6–8 and G9–12 Pair 1 and Pair 2 items. The color overlay and color contrast tools were significantly more frequently activated in the TEI version of the G2–3 Pair 1 items and the G9–12 Pair 2 items. Although no significant results were found in the use of the help tools for any item, it is worth noting the unexpectedly high activation of these tools for one TEI in particular: the TEI version in the G9–12 Pair 2 items.

The TEIs at G9–12, particularly the one in Pair 2, elicited a notably high amount of tool activation, while being more difficult and time-consuming—yet also less discriminating than their MCI counterparts. The TEIs at G9–12 also had the highest item durations and lowest efficiency of any TEI, which indicates the possibility of unique design aspects that slow and confuse test takers. Both item pairs for G9–12 were related to selecting a college. For Pair 1, test takers were given a map and a short text prompt, and they responded to a question about where students can catch a train to a college campus. In the MCI version, the prompt and map in each contained sufficient information to answer the question, and answer choices were one-word text responses. In the TEI version, the map was converted to an answer area with hotspot choices. The choices were unlabeled, so some information in the MCI was lost in the TEI. For Pair 2, the TEI was similar to the traditional MCI, but the item interface included sentences that test takers could click and drag to insert in a paragraph. The draggable options resembled clickable choices, and the confusion this might have caused could be the root of the increased tool use. In this pair, the TEI required both additional processing and interpretation of item content and response processes, increasing item duration and reducing the clarity and engagement of the items. The response area provided more of an obstacle than an alternative source of information, reducing the means of information representation. The addition of cumbersome information and distractions in the response method might be at the root of increased tool activation.

Table 9. Total number of universal tool activations for TEIs vs. MCIs with chi-square comparisons.

Grade	Item Set	Type *	Help (General)	Help (Tools)	High-lighter	Magnifier	Line Guide	Color Overlay	Color Contrast	Total Tool Use
1	Pair 1	HS	4	4	305	746	59	4	5	1127
		MC	8	4	297	770	52	1	5	1137
	Pair 2	HS	1	4	168	454	34	5	1	667
		MC	1	0	129	441	43	0	2	616
	Pair 3	DD	3	1	243**	494	48	4	4	797
		MC	3	1	146	535	53	4	5	747
2–3	Pair 1	HS	20	17	3536*	2003	522	62**	66**	6226**
		MC	9	5	2339	2081	500	26	28	4988
	Pair 2	HS	7	2	334*	62	52**	8	5	470**
		MC	0	0	76	66	15	0	0	157
	Pair 1	DD	1	1	114**	50	21	2	2	1918**
		MC	3	2	37	37	1	1	0	81
4–5	Pair 2	DD	5	1	668**	239	38	8	14	973**
		MC	2	0	269	254	36	1	1	563
	Pair 3	DD	7	4	1012	341**	91	17	14	1486
		MC	2	1	989	253	65	3	3	1316
	Pair 4	HS	48	26	7482*	1563**	655**	83	90	9947**
		MC	24	13	4057	1265	400	56	59	5874
	Pair 5	HS	19	9	4811*	1433	504	69	71	6916**
		MC	7	14	3743	1427	531	43	58	5823
	Pair 1	DD	24	16	1385*	484	204**	22	22	2157**
		MC	4	3	985	457	81	23	23	1576
6–8	Pair 2	HS	13	11	4982*	1106	323**	61	73	6569**
		MC	7	3	3313	1060	233	54	59	4729
	Pair 3	HS	14	14	5954*	763	205	75	85	7110**
		MC	20	25	3868	694	218	58	82	4965
	Pair 1	HS	15	9	1842*	392	223**	16	13	2510**
		MC	2	0	1024	310	59	17	11	1423
9–12	Pair 2	DD	248***	80***	6583*	805**	1265*	68**	64**	9113**
		MC	3	2	3508	380	137	20	20	4070

*Note: HS = hotspot item; MC = multiple-choice item; DD = drag-and-drop item

** Chi-square test significant at the $p < .001$ threshold.

*** Although no test could be run due to less than five activations in the MCI condition, it is worth noting the unexpectedly high activation of the help tools in the TEI condition.

5. Discussion

The current study extends research on technology-enhanced items in K–12 ELP assessments. Previous research has investigated the effectiveness of innovative TEIs in educational testing of math and science, and it was limited to comparing traditional and innovative items with substantive differences in content and presentation. This study adds to the body of research by examining the reading test of a large-scale K–12 ELP assessment, which included TEIs content-matched with traditional MCIs. The findings detailed in this report add to our understanding of the impact even minimal technology enhancements can have on item parameters (difficulty, discrimination), item efficiency, and item accessibility. Each of these findings will be briefly summarized, followed by synthesis and recommendations for test development. Taken together, these findings are fairly consistent with previous research on TEIs, and they point to new areas of potential research on technology enhancements in computer-mediated language testing.

Regarding item difficulty (research question 1), the TEIs analyzed in the current study were found to be overall more difficult than traditional MCIs. These findings are similar to those of previous testing research on adult computer skills testing (Jodoin, 2003), nursing tests (Qian et al., 2017), and K–12 math and science assessments (Masters & Gushta, 2018; Wan & Henly, 2012). The current study adds to the body of literature by finding that TEIs in a K–12 ELP reading test were, on average, more difficult than content-matched MCIs. Because the TEIs were matched in content with MCIs and did not involve more production than their traditional counterparts, the increased difficulty was likely due to the novel format and presentation of the items. The novel functionality and interactive nature of the items may have increased item difficulty due to test takers' unfamiliarity with the item formats. As data from this study come

from field-test items in the first year of ACCESS Online administration, it may also be that observed differences were exacerbated by the overall novelty of the online testing platform.

Additionally, the differences in the difficulty of TEIs and MCIs were not consistent across grade-level clusters. TEIs were easier than MCIs at G2–3. However, for all other grade-level clusters, TEIs were more difficult than MCIs, and the difference in difficulty widened with increased grade level. One possible explanation based on qualitative analysis of the item design is that the TEIs at higher levels became more cognitively challenging along with the lengthier reading passages and more challenging content.

Similar to the findings on TEIs adding more item difficulty, findings from previous research showed that TEIs often better discriminate between low and high performance levels than do traditional MCIs. For example, Jodoin (2003) and Wan and Henly (2012) found such evidence by examining productive, non-dichotomously scored TEIs on computer skills and K–12 science tests, respectively. In the current study, which used only dichotomously scored TEIs, the relative discrimination of TEIs varied with grade level, with TEIs being marginally more informative than MCIs at G4–5 and G6–8. Based on qualitative findings, the items designed at these levels, again, likely involved more cognitive demand, but in a way which related to the measured construct. This adherence to the intended construct may account for the TEI versions' increased discrimination and relation to the overall test score. At G9–12, where the TEIs were slightly less discriminating than the MCIs, the TEIs also involved greater cognitive demand, but the additional demand may not have been relevant for language ability. An important finding from this study, however, is that when comparing the difficulty and discrimination of items across grade levels, TEIs became most discriminating at higher levels of performance at higher grade levels. Thus, TEIs might be most useful in tests for upper-level grade students, where

familiarity with computer interface functioning can be assumed. For test developers, this means that caution should be taken in employing TEIs with younger populations, as increased interface demands can artificially boost the difficulty of an item. Ensuring that technological enhancements to items provide clear goals and means of engagement with the item is key for item design for lower grade levels. Further studies should examine the differences between item performance across grade levels and analyze a larger item set.

Regarding the efficiency of TEIs (research question 2), similar to findings in previous studies (Jodoin, 2003; Qian et al., 2017; Wen and Henly, 2012), findings in this study showed that TEIs took a significantly longer time to complete than traditional MCIs across grade levels. However, the effect size of this difference between most pairs was small. As previous studies have asserted, the increased time spent on TEIs was likely due to the need to figure out the novel interfaces. As the item pairs in this study differed only in the response interface, and did not differ in content, scoring method, or response product, it is interesting to note that the increase in average item duration across TEIs due to interface alone was significant but with a small effect size. Thus, the raw impact of technological enhancements on item efficiency could be interpreted as minimal.

Increased item duration, in and of itself, is not a negative feature, especially if the cognitive demands and complexity of the item contributes to the informativeness of the item. Therefore, comparing item duration to item information to calculate item efficiency is important. Previous research (Jodoin, 2003; Wan & Henly, 2012), which looked at open-ended TEIs that were potentially scored for more points than a given traditional MCI, found TEIs to be more efficient than MCIs due to their higher information, even when they took longer to complete. However, in the current study, the added duration outweighed the discriminative power of TEIs,

in that the items were generally less efficient than MCIs. TEIs were, on average, more efficient than traditional MCIs only at G6–8, and at other grade-level clusters, few TEIs were more efficient than the MCI version. Again, the differences in efficiency were often minimal, with larger differences in efficiency between item pairs than within pairs. Altogether, these findings indicate that technology enhancements alone do not meaningfully increase or influence item efficiency. Further research with real-time test taking data (i.e., cognitive labs) can shine a light on whether the increased duration of TEIs is merely due to the time cost of examining embedded options in hotspot items and performing the drag-and-drop action in drag-and-drop items, or the response action causes noticeable confusion in test takers' approach to items. As TEIs were not found in this study to offset duration with higher efficiency, care should be taken in test design and development to remove TEI aspects which put an additional time burden on test takers.

Regarding item accessibility (research question 3), TEIs elicited more universal tool activation than did traditional MCIs. The most activated tool was the highlighter. A few TEIs elicited more use of the line guide and magnifier tools. However, test takers did not activate the help tools and color tools, which address the interface more than content, significantly more in TEIs than in MCIs. These findings echo those of Kim et al. (2019) regarding ELs' universal tool activation in ACCESS Online. Their study also found the highlighter to be the most frequently activated tool in the reading test. The current study extends this finding, as we see technology-enhanced reading items elicit even more highlighter activation than the traditional items.

With minimal research connecting technology enhancements to universal tool activation and the sporadic nature of increased tool use in TEIs in this study, we cannot yet see any clear impact of TEI interfaces on item accessibility. As the highlighter tool can be used as a memory aid, it could be that increased activation was due to the increased cognitive load of novel TEI

interfaces. However, this speculation must be substantiated by further research. Future research could also investigate the exact reasons for tool activation while completing TEIs to understand why the tools are used more often. Data to support such an analysis could come from real-time sources such as eye-tracking or think aloud protocols, and could provide more information about TEI-related universal tool use than telemetry data alone.

For each research question, via qualitative analysis, the design aspects of select items exhibiting within-pair differences in difficulty, discrimination, duration, efficiency, and tool activation were highlighted. We were able to identify how presentation of content and item response options, which differed between TEIs and traditional MCIs, could impact the clarity and representation of information of the items. Although the findings are exploratory, they provide insight into the item design features that might affect the usefulness of TEIs in the ACCESS Online reading test. Hotspot and drag-and-drop interfaces can provide a more authentic testing environment. However, beyond adding these benefits, it is critical that TEIs enhance the performance and efficiency of the test. From these qualitative findings, TEIs which condense redundant information and reduce distractions increase item discrimination. In addition, TEIs which add more means of gathering information to provide a response to items, through added visuals or interactivity, reduce the potential difficulty of the item. However, TEIs which add obstruction and reduce clarity increase in difficulty and item duration and decrease in discriminative power. These findings reinforce the implications of the quantitative results, suggesting that specifications for TEIs must focus on setting clear response actions. To increase clarity for test takers, additional awareness raising tactics can be employed in TEIs, such as animations or pulsing highlights to indicate answer choices. Additionally, care should be taken not to sacrifice helpful textual or visual information to accommodate the enhanced input format.

Findings in this study are similar to those in Russell and Moncaleano (2019), in which drag-and-drop items had the potential to lack real-world response-process fidelity. While these TEI response processes can be innovative, they fail to truly capture the authentic processes related to the target construct of an item. As in the items in G9–12 Pair 2, the drag-and-drop interface can obscure the goal of the item and create an inauthentic text construction scenario. Again, these findings are exploratory, but they serve as a guide for further, more systematic investigation of the influence of item design on item performance, particularly in regard to technology-enhanced reading items.

6. Conclusion

This report presents findings from a study that compares TEIs to traditional MCIs embedded in ACCESS Online. Similar to findings from previous research (Crabtree, 2016; Jodoin, 2003; Qian et al, 2017; Wan & Henly, 2012), these findings show that technology enhancements led to more difficult and time-consuming items, but this did not, on the whole, impact the discrimination of the items. Unlike previous studies, the TEIs in this study were similar to MCIs in terms of content, and they differed only in terms of how test takers selected responses and aspects of visual presentation of item material. This scenario allowed our interpretation of the effect of technology enhancements to relate specifically to the influence of test interface.

Study results suggest that TEIs provide a novel means of expression for test takers, which is a pillar of Universal Design for Learning (CAST, 2011). However, the functionality of TEIs tends to increase item difficulty and duration, while the benefit of TEIs in terms of additional discrimination varied among grade levels. TEIs elicited more universal tool activation than their traditional MCI counterparts, but test takers' tool use was largely limited to the highlighter. In

sum, the TEIs seem to perform similarly enough to traditional MCIs to warrant their use on computer-based reading tests without undue threats to test reliability, and they offer additional means of engagement for test takers. They perform best when representations of information in content-matched traditional items are left intact and redundant information is removed. Due to the novelty and complexity of the interface, TEIs may be more effective in upper grades, especially grades 4–8, as discrimination between ability levels was greatest in these grades.

The differences in difficulty and informativeness between TEIs and MCIs could be due fundamentally to the novel nature of TEIs or to specific features of item design. Although TEIs were, on average, more difficult and time-consuming than traditional items, differences in item difficulty, information, efficiency, and universal tool use were greater between item pairs than within item pairs. This finding will require further analysis, as we did not perform multilevel modeling, which could be used to compare within and between item pairs, to conclusively confirm that item content is more influential on item performance than technology enhancements. However, the initial results from this report suggest aspects of TEI design can mitigate the negative effects of technology enhancements.

The current study is limited in several ways. As mentioned, more specific statistical analyses that can identify interactions between individual item content and enhancements, such as factorial analyses of variance (ANOVA) or linear modeling, might be necessary to evaluate which is more influential on item performance. Additionally, we looked at only a small pool of items spread across grades 1–12. Differences between TEIs and MCIs that vary between individual grade levels should be investigated when a larger item pool is available. The items analyzed here were in their first year of ACCESS Online administration, meaning the novelty of

the hotspot and drag-and-drop interfaces was likely higher in that year than it will be in any subsequent administration.

This is the first known study to examine content-matched TEIs and MCIs for K–12 ELs. Much more research is needed on the impact of technology-enhanced items for assessing ELs. Further research may require longitudinal analysis to determine whether the novelty effect disappears after students become familiar with the item format. Future studies may also require information on test-taker cognition. It will be critical to understand why test takers spend additional time completing TEIs and what aspects of TEIs elicit the additional use of universal tools.

Finally, this study has clear implications for future item design. TEIs provide many benefits to educational assessment and test takers, including opportunities for increased use of multimedia, multiple means of providing answers and expressing knowledge, and more authentic ways to engage with a computerized testing environment. However, the potential pitfall of TEIs is that enhancements can impede the clarity of item response goals. Recommendations can be based on the quantitative findings of this study, which showed that TEIs can be more difficult and less efficient than traditional MCIs in some cases, as well as the qualitative analysis that highlighted differences between the two types of items. Care should be taken in the development of TEIs to not reduce textual or visual information when accommodating enhanced response formats. Additional efforts should be made to make hotspot and drag-and-drop options salient to the test taker. Lastly, TEIs work most effectively when they reduce content that is redundant in a traditional item. Test designers should rely on TEIs that can eliminate the need for non-interactive images as item response choices or overlap between answer choices and reading texts.

References

- Alderson, J.C. (2000). *Assessing reading*. Cambridge University Press.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US). (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Bachman, L. F., Kunnan, A., Vanniarajan, S., & Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language testing*, 5(2), 128–159.
- Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice*, 18(3), 5–12. doi: 10.1111/j.1745-3992.1999.tb00266.x
- Bennett, R. E., Morley, M., Quardt, D., Rock, D. A., Singley, M. K., Katz, I. R., & Nhouyvanisvong, A. (1999). Psychometric and cognitive functioning of an under-determined computer-based response type for quantitative reasoning. *Journal of Educational Measurement*, 36, 233–252.
- Bennett, R. E., & Rock, D. A. (1995). Generalizability, validity, and examinee perceptions of a computer-delivered Formulating-Hypotheses test. *Journal of Educational Measurement*, 32(1), 19–36. doi: 10.1111/j.1745-3984.1995.tb00454.x
- Bennet, R. E., & Sebrechts, M. M. (1997). A computer-based task for measuring the representational component of quantitative proficiency. *Journal of Educational Measurement*, 34, 64–77.

- Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research & Evaluation*, 22(1). Available online: <http://pareonline.net/getvn.asp?v=22&n=1>
- Center for Applied Linguistics (CAL). (2015). *Enhanced item types cognitive labs: Summary of findings*.
- Center for Applied Special Technology (CAST). (2011). Universal design for learning guidelines (Version 2.0). <http://www.udlcenter.org/aboutudl/udlguidelines>
- Chapelle, C. A., & Douglas, D. (2006). *Assessing language through technology*. Cambridge University Press.
- Crabtree, A. R. (2016). Psychometric properties of technology-enhanced item formats: an evaluation of construct validity and technical characteristics [Doctoral dissertation, University of Iowa]. <https://doi.org/10.17077/etd.922fbj4d>
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd Ed.). Thousand Oaks, CA: Sage.
- Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing*, 27(4), 471–491. doi:<http://dx.doi.org/10.1177/0265532209356790>
- Dolan, R. P., Goodman, J., Strain-Seymour, E., Adams, J., & Sethuraman, S. (2011). Cognitive lab evaluation of innovative items in mathematics and English/language arts assessment of elementary, middle, and high school students: Research Report. Pearson. http://www.pearsonassessments.com/hai/images/tmrs/Cognitive_Lab_Evaluation_of_Innovative_Items.pdf.

- Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics*, 27, 115–132.
doi:<http://dx.doi.org/10.1017/S0267190508070062>
- Gutierrez, S. (2009). Examining the psychometric properties of a multimedia innovative item format: Comparison of innovative and non-innovative versions of a situational judgment test. [Doctoral dissertation, James Madison University].
- Hall, T. E., Meyer, A., & Rose, D. H. (2012). Universal design for learning in the classroom: Practical applications. Guilford.
- Hao, J. & Mislevy, R. J. (2018). The evidence trace file: A data structure for virtual performance assessments informed by data analytics and evidence-centered design. *ETS Research Reports Series*, 2018(1), 1–16.
- Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice*, 20(3), 16–25. doi: 10.1111/j.1745-3992.2001.tb00066.x
- In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, 26(2), 219–244. doi:<http://dx.doi.org/10.1177/0265532208101006>
- Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, 25, 228–242. doi:<http://dx.doi.org/10.1017/S0267190505000127>
- Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40(1), 1–15.

- Katalayi, G. B., & Sivasubramaniam, S. (2013). Careful reading versus expeditious reading: Investigating the construct validity of a multiple-choice reading test. *Theory and Practice in Language Studies*, 3(6), 877–884. doi:http://dx.doi.org/10.4304/tpls.3.6.877-884
- Katz, I. R., LaMar, M. M., Spain, R., Zapata-Rivera, J. D., Baird, J. A., & Greiff, S. (2017). Validity CHAPTER 18–Issues and concerns for technology-based performance assessments. In R. Sottolare, A. Graesser, X. Hu, & G. Goodwin (Eds.), *Design recommendations for intelligent tutoring system - volume 5: Assessment methods*, (Vol. 5, pp. 209–224). U.S. Army Research Laboratory.
- Kim, A., Yumsek, M., Chapman, M. & Cook, H. G. (2019). *Investigating K-12 English learners' use of universal tools embedded in online language assessments* (WIDA Technical Report No. TR-2019-2). Board of Regents of the University of Wisconsin System.
- Lim, H. (2019). Test format effects: A componential approach to second language reading. *Language Testing in Asia*, 9(1), 1–22. doi:http://dx.doi.org/10.1186/s40468-019-0082-y
- Liu, I. F., & Ko, H. W. (2019). Roles of paper-based reading ability and ICT-related skills in online reading performance. *Reading and Writing*, 32(4), 1037–1059.
- Liu, T. S.-W., Liu, Y.-T., & Chen, C.-Y. D. (2019). Meaningfulness is in the eye of the reader: eye-tracking insights of L2 learners reading e-books and their pedagogical implications. *Interactive Learning Environments*, 27(2), 181–199.
<https://doi.org/10.1080/10494820.2018.1451901>
- Masters, J., & Gushta, M. (2018, April). *Using technology-enhanced items to measure fourth grade geometry knowledge* [Paper presentation]. Annual meeting of the American Educational Research Association, Washington, D.C., United States.

- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.
- Moon, J. A., Keehner, M. and Katz, I. R. (2019). Affordances of item formats and their effects on test-taker cognition under uncertainty. *Educational Measurement: Issues and Practice*, 38, 54–62. doi:[10.1111/emip.12229](https://doi.org/10.1111/emip.12229)
- National Center on Universal Design for Learning. (2012). UDL in your state.
<http://www.udlcenter.org/advocacy/state>
- Pae, T. (2018). Effects of task type and L2 proficiency on the relationship between L1 and L2 in reading and writing: An SEM approach. *Studies in Second Language Acquisition*, 40(1), 63–90. doi:<http://dx.doi.org/10.1017/S0272263116000462>
- Papadima-Sophocleous, S. (2008). A hybrid of a CBT- and a CAT-based new English placement test online (NEPTON). *CALICO Journal*, 25(2), 276.
<https://search.proquest.com/docview/750618000?accountid=11226>
- Parshall, C. G. (1999, February). *Audio CBTs: Measuring more through the use of speech and non-speech sound* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Montreal, QC, Canada.
- Parshall, C., Davey, T., & Pashley, P. (2000). Innovative item types for computerized testing. In W. van der Linden & G. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 129–148). Springer Netherlands.
- Parshall, C. G., & Harmes, J. C. (2014). Improving the quality of innovative item types: Four tasks for design and development. *Journal of Applied Testing Technology*, 10(1), 1–20.

- Parshall, C., Harmes, J. C., Davey, T., & Pashley, P. (2010). Innovative Items for Computerized Testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 215–230). Kluwer Academic Publishers.
- Qian, H., Woo, A., & Kim, D. (2017). Exploring the psychometric properties of innovative items in computerized adaptive testing. *Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring From an Interdisciplinary Perspective*, 95.
- Rasskazova, T., Muzafarova, A., Daminova, J., & Okhotnikova, A. (2017). *Computerised language assessment: Limitations and opportunities*. "Carol I" National Defence University. doi:<http://dx.doi.org/10.12753/2066-026X-17-110>
- Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, 23(4), 441–474. <https://doi.org/10.1191/0265532206lt337oa>
- Russell, M. (2016). A Framework for Examining the Utility of Technology-Enhanced Items. *Journal of Applied Testing Technology*, 17(1), 20–32.
- Russell, M., & Moncaleano, S. (2019). Examining the use and construct fidelity of technology-enhanced items employed by K-12 testing programs. *Educational Assessment*, 24(4), 286–304.
- Scalise, K. (2012, May). Using technology to assess hard-to-measure constructs in the Common Core State Standards and to expand accessibility. *ETS Invitational Research Symposium on Technology Enhanced Assessments*. The Center for K–12 Assessment & Performance Management. <https://www.ets.org/Media/Research/pdf/session1-scalise-paper-2012.pdf>

- Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment*, 4(6).
- Schechinger, H. (2012, April). *Innovative computerized test items: A review* [Poster presentation]. Annual University of Kansas Professionals for Disability & School of Education Graduate Student Research Conference, Lawrence, KS, United States.
- Shaftel, J. (2015). *Accessibility for technology-enhanced items* [Conference session]. Technical Issues in Large Scale Assessment Conference, Austin, TX, United States.
- Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In T. M. Haladyna & S. M. Downing (Eds.), *Handbook of test development*, (pp. 329–347). Routledge.
- Smarter Balanced Assessment Consortium (2012). Technology-enhanced items guidelines. Measured Progress/ETS Collaborative.
- Thomas, A. (2016). Evaluating the validity of technology-enhanced educational assessment items and tasks: An empirical approach to studying item features and scoring rubrics. City University of New York (Order No. 10123821). Available from ProQuest Central; ProQuest Dissertations & Theses A&I; Social Science Premium Collection. (1807128801). Retrieved from <https://search.proquest.com/docview/1807128801?accountid=11226>
- Thurlow, M., Lazarus, S.S., Albus, D., & Hodgson, J. (2010). *Computer-based testing: Practices and considerations* (Synthesis report 78). University of Minnesota. National Center on Educational Outcomes.

- Wan, L. & Henly, G.A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education*, 25(1), 58–78. doi: 10.1080/08957347.2012.635507
- Willner, L. S., & Monroe, M. (2016). *The WIDA accessibility and accommodations framework: Considerations influencing the framework development*. Board of Regents of the University of Wisconsin System.
- Woo, A., Kim, D., & Qian, H. (2014, November). *Exploring the psychometric properties of innovative items in CAT* [Paper presentation]. 2014 MARCES Conference: Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring from an Interdisciplinary Perspective. National Council of State Boards of Nursing. University of Maryland College Park, MD, United States.



Technical Report

No. TR-2020-3
December 2020

Wisconsin Center for Education Research
University of Wisconsin–Madison
1025 West Johnson St., MD #23
Madison, WI 53706

Client Services Center toll free:
(866) 276-7735

help@wida.us
wida.wisc.edu