acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

NAPLAN Online Research and Development

Report 1: Device Effect Study – Literature Review
Report 2: Device Effect Study – Field Trial

2015

NAP NATIONAL ASSESSMENT PROGRAM

NAPLAN ONLINE
RESEARCH AND DEVELOPMENT

Device Effect Study

The device effect study was conducted to ensure that NAPLAN can be equitably administered on across all devices that meet the minimum technical requirements.

The report reveals that:

- NAPLAN online is capable of being taken on a range of devices (laptops and tablets), without device effects across content domains, item types and year levels
- the key factor influencing students' interaction with online items and tests is their familiarity with the device that they are using to complete the tests
- an external keyboard is not necessary for successful interaction with online items when students are responding to tests on tablets, although external keyboards for tablets might be useful for students who are familiar with the combined use of an external keyboard and a tablet.

There were some limited device effects, which were small, not pervasive, and centred on specific item types and features. It is anticipated that as student familiarity with devices improves between now and NAPLAN going online, these minor issues will be addressed.

Results from the device effect study informed the minimum technical specifications, which are available on the ACARA NAP website.

The study was conducted in two phases: a literature review and a field trial. Reports or extracts from reports from two contractors have been provided in this document:

Laurie Laughlin Davis, Ph.D., *Device Effects in Online Assessment: A Literature Review for ACARA,* September 2015.

Laurie Laughlin Davis, Ph.D., Irene Janiszewska, Robert Schwartz, Ph.D., Laura Holland, *NAPLAN Device Effects Study*. Pearson, Melbourne, March 2016.

Device Effects in Online Assessment:
A Literature Review for ACARA

Laurie Laughlin Davis, Ph.D.
Director, Solutions Implementation
Pearson Assessment Centre

September 2015

About Pearson

As a learning company with a global reach, Pearson offers unique insights and access to research, networks, and innovations in education. Pearson’s breadth of experience enables us to build personalized and measurable learning, integrated digital content, assessment, analysis, instruction and educator development that underpin real improvements in learning outcomes for all learners. The Pearson Assessment Centre's vision is to help the world’s learners reach their educational aspirations through meaningful feedback. Our mission is to become the centre of excellence for assessment solutions worldwide. For more information about the Pearson’s assessment products and services, visit http://www.pearsonassessments.com/.

Table of Contents

Overview

The Australian Curriculum, Assessment and Reporting Authority (ACARA) is responsible for the development of the National Assessment Program in Literacy and Numeracy (NAPLAN) in Reading, Writing, Language Conventions and Numeracy. NAPLAN tests will be delivered online beginning in 2017 with all students online by 2019. Beginning in 2012, ACARA began conducting the NAPLAN online development studies to evaluate the feasibility of delivering NAPLAN tests online. ACARA has additionally developed a comprehensive research agenda which includes an evaluation of factors that might influence the comparability of student scores across digital devices. The main purpose of the 2015 device effect study is to investigate the magnitude of measurement invariance when NAPLAN is delivered across devices and to identify whether there are variances specific to domains. The study will include a range of devices in approximately 36 primary and 36 secondary schools across Australia. In addition to quantitative item performance data, systematic observation and structured interview methods will be used to collect qualitative data on the interaction and engagement of students with online tests delivered on different devices.

The purpose of this literature review is to aid ACARA in understanding the impact of testing device (especially laptops and tablets) on performance of items and students in online assessments. A further purpose of the literature review is to guide and support development of the protocols for the systematic observations and structured interviews.

Mode and Device Effects in Educational Assessment

Administration of educational assessments via digital platforms such as computers or tablets is highly desirable to allow for a richer assessment of the construct of interest than can be achieved through traditional paper based administrations. Additionally, online assessment is increasingly more aligned with the method of instruction students are receiving in classrooms. According to survey results from a 2011 report by Simba Information (Raugust, 2011), 75% of educators said students in their school districts use device technology (including tablets, smart phones, eReaders, and even MP3 players) for educational purposes in school. The report lists increased student engagement as the primary driver for incorporation of device technology and, based on observations from early school pilots, suggests that using device technology may result in higher rates of homework completion and even increased test scores. Many school systems are experimenting with Bring Your Own Device (BYOD; Johnson, 2012; Ballagas, Rohs, Sheridan, & Borchers, 2004) initiatives or rerouting technology budgets to low cost device purchases to capitalize on these perceived positive effects as well as the potential cost savings. Consistent with this increased classroom presence, the number of instructional programs and educational "apps" for devices has also continued to grow (Ash, 2013; Glader, 2013; Rosin, 2013; van Mantgem, 2008).

However, a transition to digital assessment is not without challenges. The need to offer both computerized and paper-and-pencil versions of the same measures has been especially persistent in large-scale K-12 testing because of unevenness in available technology at the local school level (Bennett, 2003; Way & McClarty, 2012). Comparing scores and intended inferences across **modes** of administration has been shown to be challenging and is becoming even more complex as new technology-enhanced items (TEIs) are introduced into online versions of tests (but not the corresponding paper versions), and as laptops and tablets become alternate input **devices** for assessment delivery. The need for researching score comparability when an assessment is delivered via both paper and computer is addressed by a number of professional bodies such as the American Psychological Association and is also covered in the Standards for Educational and Psychological Measurement (APA, 1986; AERA, APA, NCME, 2014, Standards 9.7 & 9.9). Score comparability research over the last quarter century has largely focused on differences between paper-based and computer-delivered assessment assessments (see for example: Winter, 2010; Kingston, 2009; Texas Education Agency, 2008; Wang, Jiao, Young, Brooks, & Olson, 2008; Wang 2004; Wang, Jiao, Young, Brooks, & Olson, 2007; Meade & Drasgow, 1993), while less research has mined the wider implications of Randy Bennett's definition of comparability as "the commonality of score meaning across testing conditions including delivery modes, computer platforms, and scoring presentation" (Bennett, 2003). However, with a shift towards online testing coinciding with the proliferation of devices appearing in the classroom, the sub-genre of "device" comparability is of increasing interest to assessment developers.

As the pace of technological change increases, a host of issues threaten the standardization of test administration conditions, and therefore, the comparability of test scores from those administrations. On the one hand, technology enables us to measure things in ways we have not been able to previously and may, in fact, improve measurement accuracy and validity in some cases. On the other hand, the realities of updating school technology infrastructures to keep up with changing technologies almost guarantees that there will be a perpetual state of technological differences across students, schools, and states. While it is not always clear how this device diversity should best be managed, Kolen (1999) categorizes the potential sources of mode (and, by extension, device) effect into four broad categories: test questions, test scoring, testing conditions and examinee groups. Each of these requires consideration when designing research studies to look at mode and device comparability.

Key Definitions

As we move from a general look at comparability to the specifics of comparability across devices, it is worth pausing to provide some definitions. **Device** in the broadest sense encapsulates a range of technology hardware used to access digital content and can include a wide array of input mechanisms, output mechanisms, shapes and sizes. Examples of digital devices include iPods (which access audio content), eReaders (which primarily access text content), smart phones , tablets, laptop computers, and even desktop computers.

As a digital device, a **tablet** is usually larger than a smart phone but differs from a laptop or desktop computer in that it has a flat touch-screen and is operable without peripherals like a mouse or keyboard. In regards to size, most tablets weigh 1 to 2 pounds and are designed to be handheld, used in a flat position, placed in a docking station, or held upright by a foldable case. A tablet has no singular correct position, which is reinforced by re-orientation of the on-screen image to portrait or landscape based on the position of the device. Although the most typical tablet screen size of 10 inches was popularized by Apple's iPad, smaller screen sizes of 5, 7, or 8 inches are gaining in popularity. For instance, the 7-inch Samsung Galaxy Tab, with about half of the surface area of a 10-inch device, resembles the size of a paperback book. Apple has also recently introduced the iPad mini which is a smaller (8-inch) version of the full size iPad.

The boundaries between tablets and computers sometimes blur with quick-start computers and with hybrids and convertible computers, which combine touch-screens with keyboards that can be removed, swiveled, or tucked away. In another instance of blurring boundaries, some e-readers are becoming increasingly indistinguishable from tablets in their use of mobile operating systems, similar size and shape, color touch-screen, long battery life, wi-fi connectivity, and support of downloadable “apps.” However, e-reader screens, unlike tablet LCD screens, are optimized for reading even under bright light conditions, while tablets tend to be designed with more memory and storage space for supporting multiple media forms and a wider range of applications. Additionally, smart phones are becoming larger and more similar to small tablets reflecting what a new blended category called a “phablet.”

The degree of comparability that might be expected of scores from different devices may depend to a large extent on the differences in device **form factor** (i.e., size, style and shape as well as the layout and position of major functional components) and how students interact with the devices themselves (Way, Davis, Keng, & Strain-Seymour, in press). For example, as supported by previous research with students taking the Graduate Records Exam (GRE; Powers & Potenza, 1996), comparability between desktops and laptops can be expected to be relatively high because the form factors of the devices are fairly similar. Both desktops and laptops have physical keyboards (though they may vary in size and key positioning and spacing) which are positioned horizontally on a flat surface relative to a screen or monitor which is positioned vertically (though they may vary in size and flexibility in the degree of vertical placement). Similarly, both are used with a mouse as the pointer input for selection (though they may vary in terms of whether the mouse pointer is controlled through an external mouse or a touch pad). Conversely, comparability between desktops and smart phones can be expected to be lower as the form factors of the devices are relatively dissimilar. Unlike a desktop, a smart phone typically has an onscreen keyboard which is overlaid on top of the small screen (typically 4-5 inches) when opened, expects touch input for selection, and may be flexibly used in a variety of hand-held positions and distances for viewing relative to the face. Thinking through how the form factor of a device influences how information is viewed and accessed by the student as well as how the student provides a response to a question is the logical first step in any evaluation of cross device comparability.

Assessment Conditions by Device Type

Most assessment programs have adopted a set of hardware and software specifications for use in delivering online assessment. For computers, this might include the minimum processor speed, amount of RAM memory, monitor size, etc. For tablets, this might additionally include minimum screen size requirements, external keyboard requirements, operating systems, etc. These specifications are often developed to support the optimal delivery of the online test, but should also account for conditions likely to promote comparability of students’ experience during testing. For example, in US the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced assessments programs require tablets to have a minimum screen size of 9.5” and use an external keyboard (PARCC, 2013; SBAC, 2013). These requirements allow a range of different tablets and operating systems to be used but hold them all to the same expectations relative to size and text input which allows a greater degree of standardization in the test administration. The US based National Assessment of Educational Progress (NAEP) has taken this degree of standardization a step further. NAEP has announced plans to begin digital delivery of assessments within its program in 2017 will use a standard set of tablets that are provided to schools for use during the NAEP administration. In this way, NAEP explicitly controls for the device on which students test (McCullough, 2015).

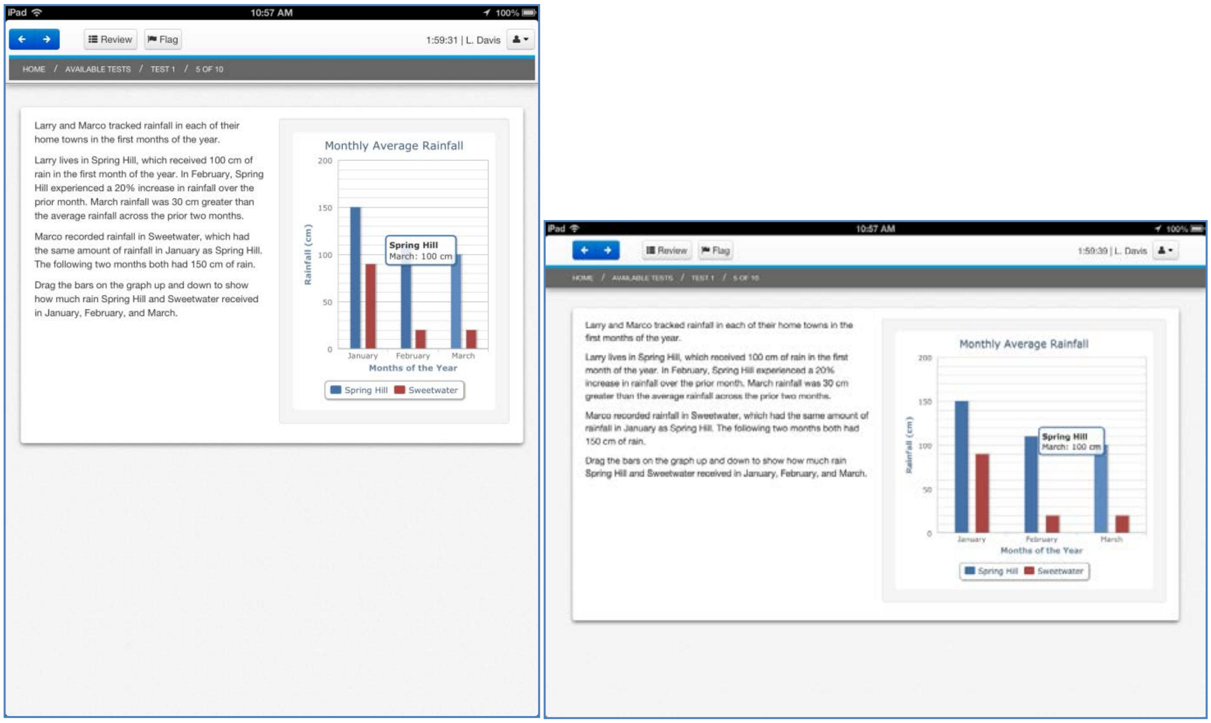
Touch-Screen Specific Features

Touch-screen devices allow for certain interactions and experiences not available with computers. For example, pinch-and-zoom magnification, screen rotation (landscape to portrait), and autocorrect or auto-complete are all features common to tablets and smart phones, but not frequently seen with computers. While not an inherent property of the touch interface, the purpose of these features is to offer alternative interactions to compensate for certain limitations of the device size and input mechanisms. The challenge for comparability occurs either when the features advantage or disadvantage users of the touch devices or when the features violate the measurement construct (Way, Davis, Keng & Strain-Seymour, in press).

Screen Orientation

Screen rotation from landscape to portrait is typically considered a positive attribute for tablet and smart phone applications. However, in considering issues of device comparability between tablets and computers in assessment settings, the differences inherent in how test content is displayed and viewed in a portrait versus a landscape orientation may create challenges (Way, Davis, Keng & Strain-Seymour, in press). Computer monitors are typically (though not always) landscape orientation. Test content which is designed for computer delivery may not translate well to a portrait orientation on a tablet (see Figure 1) as scrolling might be introduced and long horizontal elements (like rulers or number lines) may not scale well. For this reason, testing programs may be better served by disabling screen rotation when presenting test items on tablets.

In a study of elementary school students (grades 4-6) and middle/high school students (grades 8-11), Davis (2013) explored student preferences around device orientation when viewing assessment content by structuring research conditions where students were handed the tablets by study facilitators either in portrait orientation or in landscape orientation without specific instruction of what orientation should be used. All students who were handed the tablet in landscape orientation kept that orientation during their testing. Of students who were handed the tablet in portrait orientation, only one student rotated it to landscape orientation (though a second student rotated the tablet 180 degrees and kept it in portrait). If prompted at the end of the session, students acknowledged that they knew they could rotate the tablet, but did not feel the need to do so. This suggests that students were able to comfortably read the questions and the passages even in the narrower portrait orientation. While the use of portrait orientation of the tablet did not seem to adversely affect students in this study, neither did it seem to add value to the test taking experience. Some students could conceive of using the portrait orientation for something like an eReader, but would expect different behaviors in portrait mode to support this (like full screen capability for the reading passage). Additionally, students within this study (and previous studies) generally did not attempt to change the screen orientation from the way in which the tablet was handed to them, suggesting that this is not a highly valued interaction by students.



Larry and Marco tracked rainfall in each of their home towns in the first months of the year.

Larry lives in Spring Hill, which received 100 cm of rain in the first month of the year. In February, Spring Hill experienced a 20% increase in rainfall over the prior month. March rainfall was 30 cm greater than the average rainfall across the prior two months.

Marco recorded rainfall in Sweetwater, which had the same amount of rainfall in January as Spring Hill. The following two months both had 150 cm of rain.

Drag the bars on the graph up and down to show how much rain Spring Hill and Sweetwater received in January, February, and March.

Larry and Marco tracked rainfall in each of their home towns in the first months of the year.

Larry lives in Spring Hill, which received 100 cm of rain in the first month of the year. In February, Spring Hill experienced a 20% increase in rainfall over the prior month. March rainfall was 30 cm greater than the average rainfall across the prior two months.

Marco recorded rainfall in Sweetwater, which had the same amount of rainfall in January as Spring Hill. The following two months both had 150 cm of rain.

Drag the bars on the graph up and down to show how much rain Spring Hill and Sweetwater received in January, February, and March.

Figure 1: Double-bar chart question shown in portrait view (left) and landscape view (right)

Pinch and Zoom

Students can use pinch-and-zoom magnification within a touch environment to enlarge portions of the screen and examine content in more detail. This might be viewed a positive attribute in terms of overcoming the smaller screen size of tablets and smart phones and may be especially valuable for improving the readability of portions of text. However, it should be recognized that when a student zooms in, he or she is no longer able to view the item content in its entirety and may have to zoom back out to view information necessary to correctly answer the item. Additionally, while enlarging text may allow for better readability of the portion of the reading selection students are viewing, it may create other challenges for reading recall as it is more difficult for a student to retain their awareness of where information is “on the page” and to use other types of visual landmarks when zooming in and out (Way, Davis, Keng, & Strain-Seymour, in press).

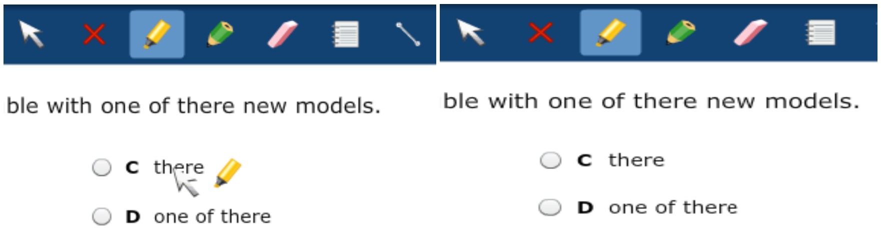
Davis (2013) found that the use of pinch and zoom was mixed across students. Most students knew they could pinch and zoom, might experiment with this on one question, and would immediately attempt this action if prompted for a way to make the text bigger. However, only a handful of students used pinch and zoom with any regularity during their test sessions. Most students were able to adequately view the test content without need for magnification. Additionally, those students who did desire magnification were familiar enough with tablets to attempt a pinch and zoom gesture. It remains unclear, however, to what degree this magnification benefits students since in the process of enlarging part of the screen content, they lose visibility of other parts of the content. Thus it is clear that this gesture does not directly act as an offset for the smaller screen size of tablets relative to larger desktop computer monitors.

The Finger as a Pointer

Within studies of input devices such as touch-screens, comparisons are made between the benefits of the immediacy of direct input, where moving on-screen objects resembles moving objects in the physical world, and those of mechanical intermediaries, such as the indirect input of a mouse. While speed, intuitiveness, and appropriateness for novices are benefits of direct input, mechanical intermediaries often extend human capability in some way (Hinckley & Wigdor, 2011). Similarly, tablets' touch input is immediate and direct, while mouse input aids accuracy and allows one small movement to equate to movement of the cursor across a much larger screen distance.

Touch inputs are associated with high speed but reduced precision; they are typically faster than mouse inputs for targets that are larger than 3.2 mm, but the minimum target sizes for touch accuracy are between 10.5 and 26 mm, much larger than mouse targets, which tend to be more limited by human sight than by cursor accuracy (Vogel & Baudisch, 2007; Hall, Cunningham Roache, & Cox, 1988; Sears & Shneiderman, 1991; Meyer, Cohen & Nilsen, 1994; Forlines, Wigdor, Shen, & Balakrishnan,2007). Touch-screen input accuracy may suffer from spurious touches from holding the device and from occlusion when the finger blocks some part of the graphical interface (Holz & Baudisch, 2010).

A mouse-controlled cursor can be moved without triggering an active selection state; cursor movement is differentiable from dragging. The cursor shows the user the precise location of the contact location before the user commits to an action via a mouse click (Buxton 1990; Sutherland 1964). A touch-screen, on the other hand, does not have these two distinct motion-sensing states; pointing and selecting, moving and dragging, are merged. No "hover" or "roll-over" states as distinct from selection states can exist on a touch-screen, which removes a commonly used avenue of user feedback within graphic user interfaces. Similarly, without a cursor, touch-screen interfaces cannot have cursor icons, which can be used to indicate state or how an object can be acted upon (See Figure2; Tilbrook 1976). For these reasons, it is important to consider touch-screens when designing user interfaces and item interactions especially when test content will be delivered across multiple device platforms.



ble with one of there new models.

○ C there

○ D one of there

ble with one of there new models.

○ C there

○ D one of there

Figure 2. Cursor icon on computer (left) showing the highlighter tool selected is absent on tablet (right)

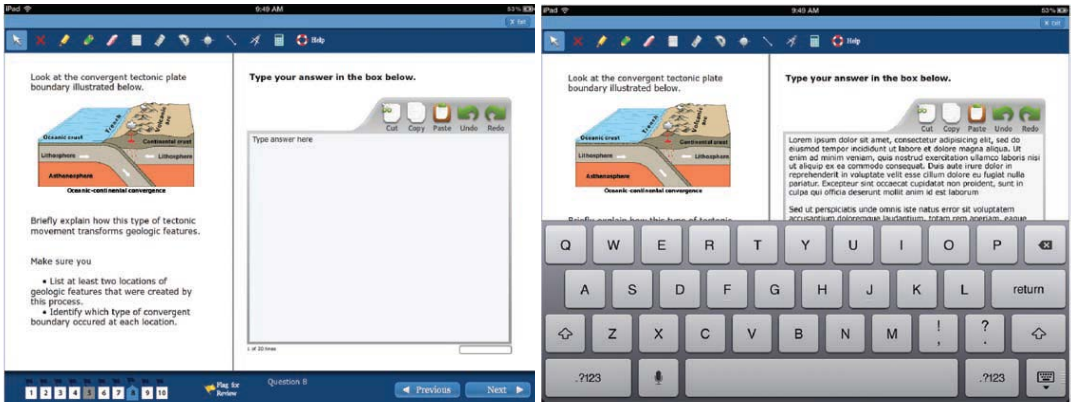
Use of Onscreen vs. External Keyboards

With regard to keyboard functioning, it is important to draw attention to the difference between an **onscreen keyboard** (sometimes also called a virtual keyboard or soft keyboard) available with touch screen devices and an **external keyboard** (sometimes also called a physical keyboard) which may be used with either touch or non-touch devices. The onscreen keyboard is a piece of software native to the device which may be launched by the user or launched automatically within certain applications. This allows the user to provide text input to the device without the need for a separate peripheral. By contrast an external keyboard is a physical peripheral that must be connected to the device either by wire or wirelessly to support text entry.

Onscreen keyboards

The onscreen keyboard differs from a traditional physical keyboard in a number of ways. Perhaps most importantly, the virtual keyboard allows two states relative to hand positioning (fingers are off the keys; fingers are depressing a key) compared to the three states possible with a physical keyboard (fingers are off the keys; fingers are resting on the keys; fingers are depressing a key) (Findlater & Wobbrock, 2012). This lack of a resting state with an onscreen keyboard creates challenges for the use of conventional keyboarding techniques where students are taught to rest their fingers on home row. The compact size of an onscreen keyboard also means that students are working in a more constrained space when reaching fingers to select keys. Both of these issues result in significantly slower typing speeds when using an onscreen keyboard (Pisacreta, 2013).

Additionally, onscreen keyboards typically have multiple screens with alpha characters displayed on one screen and numeric and symbolic characters displayed on one or more alternate screens. Students must know how to navigate between these screens and where to find the specific character for which they are looking (Davis & Strain-Seymour, 2013b). Because of this, students can make typing mistakes when using the onscreen keyboard that they would not normally make when using a physical keyboard such as selecting a comma instead of an apostrophe (Lopez & Wolf, 2013). Similarly, the onscreen keyboard is hidden when not in use which requires students to know how to open and close the keyboard when needed. When open, the onscreen keyboard takes up a significant amount of screen real estate and often pushes content off the screen requiring the student to scroll up to locate or reference information not in view (see Figure 3).



Look at the convergent tectonic plate boundary illustrated below.

Briefly explain how this type of tectonic movement transforms geologic features.

Make sure you

- List at least two locations of geologic features that were created by this process.
- Identify which type of convergent boundary occured at each location.

Type your answer in the box below.

Type answer here

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque

Figure 3. The onscreen keyboard takes up screen real estate and may cause scrolling

Several observational studies suggest that students tend to write less under these circumstances than they would with an external keyboard (Davis, Strain-Seymour, and Gay, 2013; Strain-Seymour, Craft, Davis, & Elbom, 2013). However, Davis, Orr, Kong, and Lin (2015) did not find differences in either the length or quality of student writing when using an onscreen versus an external keyboard in a study conducted with 5th grade and high school students (grades 10 and 11). Additionally, it should be noted that many students (especially younger ones who have not yet perfected keyboarding skills) expressed that they preferred the onscreen keyboard and some even theorized that it might be faster for them than an external keyboard(Davis, Strain-Seymour, and Gay, 2013).

Accommodations for Text Entry

Touch-screen device manufacturers have taken steps to enhance the usability of the onscreen keyboard with features such as haptic feedback (the key vibrates to let the student know it has been selected) or a magnified image of the key which is highlighted when the key is pressed. In fact, students report that they like these and other features of the onscreen keyboard. For example, some students report that they like that, when activated, the caps lock key on the virtual keyboard turns all the letters from lower case into capital which makes the state of the caps lock function very clear (Strain-Seymour, Craft, Davis, & Elbom, 2013).

With an onscreen keyboard, different configurations of keys are possible though the use of application settings. For example, split keyboards break the standard QWERTY keyboard interface into two sections which appear split between the lower left and lower right portions of the screen. This may facilitate text entry through more of a thumb-style texting process. Additionally, many touch screen devices allow for a gliding or dragging motion across keys (see for example www.swype.com) combined with predictive algorithms as a style for text entry. Findlater & Wobbrock (2012) discuss an adaptive onscreen keyboard which learns a users keystroke motions over time and adapts the key placement to best suit their finger length and reach.

Additional features such as autocorrect and autocomplete complement an array of keyboard configuration options. Autocorrect is similar to a spell check feature which is usually present in most word processing software packages on computer. With autocorrect enabled misspelled words on a touch-screen device may be automatically corrected without the student having to make the correction themselves. This could provide both advantages and disadvantages relative to student performance—especially if spelling is part of the construct being measured (Way, Davis, Keng, & Strain-Seymour, in press). Autocorrect could provide an advantage to touch-screen users by enabling them to spell poorly and rely on the tool to correct their mistakes. However, the autocorrect functionality may misinterpret the word the student is trying to type and correct to a different word altogether. Autocomplete goes a step beyond autocorrect and applies predictive algorithms to complete or suggest words to students based upon the first few letters typed. Given the limitations previously discussed with onscreen keyboards, this feature has some attraction relative to leveling the playing field with external keyboards. However, this feature may go too far in providing assistance to students with word choice within their academic writing (Way, Davis, Keng, & Strain-Seymour, in press). Alternatively, it may disadvantage students because it encourages them to pick words from a list without regard to their appropriateness in context.

External Keyboards

External keyboards for tablets can roughly be classified into four categories—stand-alone keyboards, keyboard shells, folio case style keyboards, and clamshell keyboards (see Figure 4; Frakes, 2013). Stand-alone keyboards are completely separate from the tablet, but generally provide an experience most similar to a full size keyboard (full size keys, more space between keys, etc.). However, they require the addition of a stand to use tablet in vertical or propped up orientation relative to the keyboard. With a keyboard shell, the keyboard is contained within a hard case that covers the tablet screen when not in use, but can also be used as a stand for the tablet. The keys are smaller than a full size keyboard and spacing tends to be more cramped than with a full size keyboard. For a folio case style keyboard, the keyboard is integrated into a soft folio style case for the tablet which can be folded to prop up the tablet when keyboard is in use. The keys are smaller than a full size keyboard and spacing tends to be more cramped than with a full size keyboard. Finally, with a clamshell keyboard, the keyboard attaches to the tablet though a hinge or locking mechanism at the rear of the tablet and functions more like a laptop. The keys are of higher quality, but spacing still tends to be more cramped than with a full size keyboard.

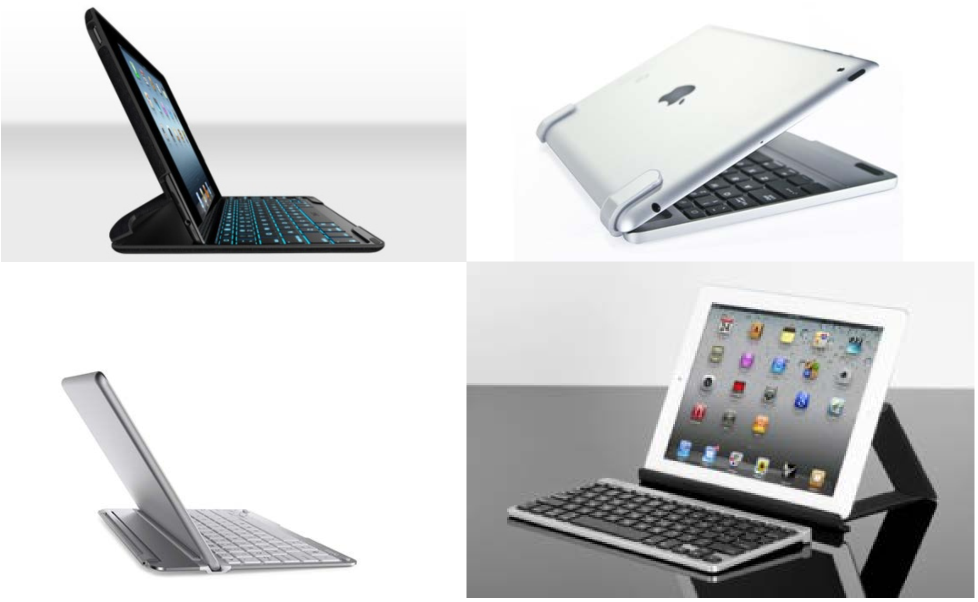


Figure 4. Different types of external keyboards for tablets. Folio case, upper right; Clamshell, upper left, keyboard shell, lower right, stand-alone keyboard, lower right

External keyboards are not necessarily the magic solution to the challenges of device comparability (Davis & Strain-Seymour, 2013b). Because of the variability in external keyboards, simply requiring their use with a tablet for assessment purposes may be insufficient to produce the intended degree of standardization across devices. Additionally, the use of an external keyboard more or less dictates the position in which a student will work with the tablet—either propped up on a case or in a stand. Davis, Strain-Seymour, and Gay (2013) reported that some students appeared to find it difficult to switch between using the external keyboard to type and using their finger to select text and place the cursor. One student characterized this drawback as “everything not being in one place.” Lastly, use of the student’s finger to place the cursor in the proper spot for editing text provides additional challenges as students may have trouble getting the cursor to the right spot.

An additional consideration with requiring the use of external keyboards for assessment is whether the students work with those keyboards regularly during other academic activities. If they do not and external keyboards are only brought out on testing day, this might introduce construct irrelevant variance into test scores. In fact, many schools do not purchase external keyboards or stands with tablets and might not have such equipment available. Recall as well that younger students—those who have not yet learned keyboarding skills and who might best reflect what Prensky (2001) termed “digital natives” might actually be more facile in entering text without the external keyboard.

The relationship of keyboards to score comparability did not originate with advent of onscreen keyboards. Some early research (Powers and Potenza, 1996) in device comparability considered differences between laptop and desktop keyboards. This study did find differences in student writing performance which they believed to be due to the different sizes and layouts of the device keyboards. Similarly, Sandene, Horkay, Bennett, Allen, Braswell, Kaplan, and Oranje (2005) reported findings from the 2002 NAEP online writing study in which there was some evidence that students testing on laptop computers had lower writing performance than students writing on desktop computers. However, this finding was not consistently observed across essay prompts or studies within the NAEP research. Additionally, the NAEP online writing studies found that online writing performance (on any device) seemed to be related to computer familiarity. As we consider keyboard options with tablets it is important to note, however, that nearly a decade after this research there seems to be little attention given to this issue as users' facility with the laptop keyboards as well as the design of the keyboards themselves has improved.

Use of Tools

Within an online testing environment, a variety of digital tools may be available for students' use. In general, these digital tools are intended to approximate functionality that students have when taking a test in the physical world such as marking with a pencil or using a calculator. Marking and note-taking tools (e.g. highlighter, pencil, etc.) might be used to chunk information within an item as part of a test-taking strategy or might be used as an approximation for hand-written calculations. Other tools are content specific for mathematics or science such as calculators, rulers, or protractors.

In designing tools for use within an online testing program, it is important to consider usability of the tools across all devices that students will use to access the test (Davis, Strain-Seymour, & Gay, 2013; Strain-Seymour, Craft, Davis, & Elbom, 2013; Yu, Lorié, & Sewall, 2014). This means considering how students will use either the mouse or their fingers to interact with each tool. For example, if a student is intended to use a mouse to precisely move and rotate a ruler into position, a student testing on a tablet should be able to move and rotate the tool with that same level of precision using only their finger. In a study with students from grades 4, 8, and high school, Strain-Seymour, Craft, Davis, & Elbom (2013) found that, in general, students working with tablets were able to use tools with ease, and a few even indicated that it was easier because it was more "direct" or because you did not need to use the mouse. However, when students tried to use the marking tools to highlight single words, circle a significant word or number, or underline a phrase, several students tried a couple of times to position the mark correctly with a couple acknowledging that the mark was not made exactly where intended but "close enough."

Additionally, consideration should be given to making sure that tools are well managed with regard to screen real estate and do not obscure testing content that might need to be referenced as they are used. For example, students should be able to position the calculator on their screen such that it does not overlap with the information in the item they need to reference to make their calculations. Consideration must also be given as to whether and how tool marks (highlights, pencil notations, etc.) are associated with an item. For example, if a student navigates to another item and back to the original item, will the original tool marks be stored and displayed? How will this be handled if a set of items is associated with a passage? Will tool marks be allowed within a passage and will they be available across all items within the passage set? Finally, how might tool marks be displayed if students are allowed to use pinch and zoom or screen rotation capabilities of a touch-screen device?

Davis, Strain-Seymour, and Gay (2013) also suggest the need to clearly differentiate for students between marks that indicate an item response and tool marks that do not. Some students in their study assumed that they needed to use the dot tool to indicate their response to a hot spot item by placing a dot on the selected response. However, since marking tools (like the dot tool) are intended to support a student's thought processes but are never scored, this was a problematic issue. Similarly, they found with graphing items some students did not realize that they could simply touch the graph to plot a point. Instead they attempted to use the dot tool or line tool to interact with the item. This issue can be avoided by not offering marking tools on items where the tool mark might be confused with the actual response mechanism. Alternatively, students might be prompted that they have not responded to an item either when navigating away from the item or prior to submitting their test.

Research by Subject Area

While the formal psychometric study of device effects across computers and tablets is still nascent, there seems to be a general trend toward non-statistically significant and/or non-practically significant effects at the total score level. For example, in a study of device comparability for the PARCC assessments for students in grades 4, 8, and 10, Keng, Davis, McBride, & Glaze (2015) concluded that "most differences between the tablet and computer conditions were small in magnitude and nonsignificant based on the statistical criteria used in the study." This is consistent with conclusions from Olsen (2014) which stated that "there is strong evidence that STAR Reading Enterprise and STAR Math Enterprise were measuring the same attribute regardless of device type." This is also consistent with research from Davis, Orr, Kong, & Lin (2015) which compared computers and tablets (with and without external keyboards) for assessment of written composition and found "no significant performance differences observed in student writing across study conditions." Additionally, in a study of reading, mathematics, and science assessment with students in grade 11, Davis, Kong, & McBride (2015) found that "there were no observable performance differences in student test scores across device for any of the three content areas." Taken together, these studies suggest that while there may be specific instances where user interface or item type might create barriers for student in engaging with assessment content from a particular device, that assessment developers can generally expect a high degree of comparability across computers and tablets. This section delves more specifically into user interface and results for key subject areas.

Reading

Displaying Items and Passages

Reading/Literacy is an important but complex area for consideration when assessing across devices as students interact not only with the items themselves, but also with reading passages. Screen real estate must be balanced between the item itself which may involve one of many different response formats (see section on item types below) and passages (which often contain information that students will need to reference in selecting or creating their response to the item). With regard to digital presentation of reading passages, two primary options have gained popularity for use with assessment programs. Passages might appear first on a screen by themselves with the items following on successive screens. In this interface option the passage may appear in a reference panel or window that can be brought up by students as they review the items but is otherwise hidden from view when students are reviewing items. Alternatively, the passage might appear side-by-side with items where as students navigate from one item to the next the passage side of the screen stays the same, but a new item is displayed.

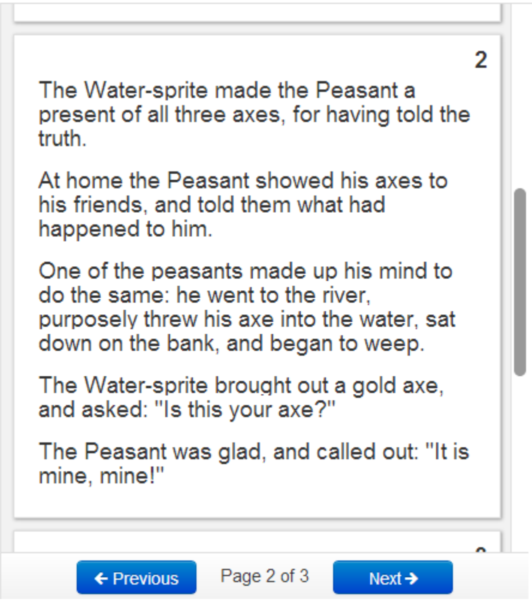
In an investigation of reading interfaces for computer and tablet, Strain-Seymour and Davis (2013) found that some younger students' had tendency to re-read the entirety or a large portion of a passage when it appeared multiple times paired with each item. The researchers suggested that one possible solution would be to use a passage introduction screen and to switch the side of the screen that the passage appears on. In this case, the passage would first appear without potential distraction from the item being in view on the same screen. The passage would then also appear on the right as students navigated through the items. The introductory text could also include an additional instruction: "Questions about this passage may appear with the passage repeated next to the question for easy reference." With the item on the left on subsequent screens, the student would most likely start by reading the question rather than feeling compelled to read the passage again.

Reading interface design becomes even more complex when the test design may ask students to read and synthesize information across multiple reading passages when responding to items. This type of "paired passage" further exacerbates the impact of limited screen real estate available for tablets. One solution is to use tabbed display of passages so that students can toggle back and forth between passages by selecting a tab at the top or bottom of the screen. The solution appears to work well for both tablets and computers so long as the tabs themselves are sufficiently large for students to select with their fingers.

Managing Scrolling with Lengthy Passages

Depending on the length of the passage students are asked to read, scrolling might result. **Scrolling** is the digital equivalent of turning a page in that students need to "scroll up and down" (or sometimes scroll left and right) to see the full text of the passage. This metaphor turning the page actually suggests one interface solution for addressing this called the "paging" interface. In this design, text is pre-sectioned into different pages with clear page breaks between them and students using a navigation style button to move from one page to the next. By contrast, in a "scrolling" interface students more fluidly scroll through text as a continuous stream without pre-determined page breaks. Scrolling is managed either through scroll bars using the mouse or finger or, in the case of a touch screen device, may be accomplished by direct manipulation of the passage with the finger. Prior research has suggested the benefit of a paging interface over a scrolling interface for supporting the relocating of content using visual memory (Higgins, Russell, & Hoffman, 2005; Pommerich, 2004). The visual memory support of a paging interface may come from the fixed position of text on screen (a given sentence always appears in the same position on screen) and/or from the position of a given sentence in relation to the top or bottom of a page (e.g., a sentence appears just before the end of the page).

Strain-Seymour and Davis (2013) evaluated a hybrid interface where students in tablet and computer conditions could either scroll or use paging buttons to move through the text (see example in Figure 5). Within this design, the contents of a single page remained constant; the same text, the same column width, and roughly the same line breaks characterize a page, even on a larger monitor where 1.5 pages, for instance, might be visible at a time. In this study, students universally recognized when passages had more text than could be seen on the screen at once. Most students could cite the presence of the scrollbar as the indicator for additional content and of the current position within the overall passage. When the passage was presented using an interface that both paged and scrolled, students used the scrolling more frequently than the paging. Paging provided an alternative when a student struggled with scrolling (tablet and computer) or tried to scroll but accidentally highlighted or extracted text within text highlighting and text extraction items (tablet only). When an attempted scroll led to an inadvertent response within a text highlighting item, most students seemed to be aware that they had accidentally changed their answer.



The Water-sprite made the Peasant a present of all three axes, for having told the truth.

At home the Peasant showed his axes to his friends, and told them what had happened to him.

One of the peasants made up his mind to do the same: he went to the river, purposely threw his axe into the water, sat down on the bank, and began to weep.

The Water-sprite brought out a gold axe, and asked: "Is this your axe?"

The Peasant was glad, and called out: "It is mine, mine!"

Figure 5. Example of a hybrid paging-scrolling interface for reading passages

The only action that was harder on the computer than on the tablet was scrolling a passage, if accomplished by grabbing the small scrollbar. Computer users who, instead of trying to grab the small scrollbar, used the scroll wheel on the mouse to scroll the passage, had no such difficulty and, of all user types, were able to scroll the passage with the greatest ease and most precision, Tablet users — who only had one way to scroll the passage — were in the middle of the difficulty spectrum. Most students on the tablet knew to swipe, with only a few initially trying to swipe in the wrong direction.

A few students used the paging and scrolling for different purposes. Some students were adept at using the scroll wheel on the mouse to scroll the passage a line at a time, thereby keeping their eyes in roughly the same position. (The same behavior did not occur as frequently or as seamlessly on the tablet where the scrolling action is not as easy to control down to the precision of a single line.) Others adjusted the scroll position to show a complete section of a passage at a time, positioning the bold heading at the top (something that is not possible in an interface that only pages). A few students who otherwise tended to scroll used the paging functionality when trying to move very quickly to an early or late part of the passage to re-read a portion.

Research Study Results for Device Effect

Despite all the consideration given to optimizing the design of a reading passage interface for assessment, there is minimal evidence for strong or consistent device effects between tablets and computers for reading assessments within quantitative studies of device comparability. Olsen (2014) reports statistically significant but very small effect sizes for reading which varied in direction across grade levels for students in grades 1-11 (tablet favored in half the cases and computer favored in the other half). Olsen recommends against making an adjustment to scoring based upon device. Keng, Davis, McBride, & Glaze (2015) report statistically significant differences in favor of computer for an integrated reading-writing assessment of English language arts for 4th grade students, but attributed this to lack of student familiarity with innovative assessment tasks as well as methodological challenges with creating a matched sample.

Davis, Kong, & McBride (2015) report no significant differences between tablet and computer conditions for reading. However, they note that despite the lack of statistically significant differences in student performance for reading, there is some evidence to suggest that reading may be an area where students prefer to work with tablets. Specifically, they observe that the scrolling interface used to present the reading passages in their study provided for a very natural gesture with the finger on the touch-screen device whereas the use of the mouse as an intermediary device to scroll the passage on the computer may be somewhat more cumbersome. Additionally as tablets and eReaders become more and more common, students may find interacting with text on a touch-screen device a very familiar activity.

Mathematics/Numeracy

Assessment of mathematics/numeracy in a digital environment allows for a broader range of item interactions than can be included in a paper-based environment. However, the fidelity of these interactions to actual math processes used in the classroom can create challenges if students have to learn the interface just to take the assessment and do not work with those types of digital interfaces regularly in instruction. For example, an article in Education Week (Heitin, 2014) raised the issue of student familiarity with solving open ended math problems through a digital interface, in general, and suggests that the tools and interfaces available for students to create their responses are artificial and do not match the largely still paper based way students work through math problems in the classroom. Math assessments also tend to make greater use of tools than do reading assessments with many tools that are construct relevant or required to solve problems (e.g. calculators, rulers, protractors, etc.). The design of these tools is as important as the design of the user interface for the items themselves. As previously discussed, the tools need to be usable for students interacting with them either via mouse or finger.

While these issues arguably apply across multiple device contexts, they might be exacerbated in situations where students' use of devices for certain interactions may be limited. Unlike reading, where students have the potential for many applications of tablets outside the classroom (for reading e-mails, magazines, eBooks, etc.) it is unlikely that many students would use tablets for math related interactions unless they were instructed to do so as part of their academic programs. This sentiment is echoed in conclusions reached by Keng, Davis, McBride, & Glaze (2015) where they suggest that the small effects observed for 4th grade students in mathematics could reflect a lack of familiarity and comfort with entering responses to mathematics tasks on tablets. Olsen (2014) also found practically small but statistically significant results at several grade spans encompassing grades 1-11 favoring the computer. While he noted that the effect sizes were not large (double the size of reading effect sizes, but still "small" or "very small") and score adjustments were not warranted he also suggested continued monitoring of score differences during operational implementation. Davis, Kong, & McBride (2015), however, found no evidence of statistically significant device effects for mathematics.

Writing

Assessment of written expression is integrally entwined with the method of text entry for tablets. Selection of keyboard and allowance for different settings and features can be expected to drive outcomes as much as the structure of the writing tasks and interfaces. Additionally, as classroom instruction in writing becomes more integrated with technology these technologies may fundamentally alter the construct being assessed. Way, Davis, & Strain-Seymour (2008) reported that the cognitive processes students use when writing depends to a large extent on what tools they are using to write. When handwriting compositions, students typically make use of "pre-writing" skills such as drafting and outlining; however, when students have word processing tools available they use fewer pre-writing skills and instead opt to write and revise on the fly using the convenient revision tools within the word processing software. Touch-screen specific tools such as autocorrect and autocomplete could arguably once again change the way students engage cognitively with the writing task. In fact, technology such as speech-to-text enables text entry without the need to "write" at all. Written composition as a method of communicating ideas for an audience could become obsolete or at a minimum become relegated to a tool for revision of an initial spoken draft. While these types of technology driven changes may never come to pass, this underscores the importance for assessment programs of defining the construct of writing relative to the use of technology supports and the alignment of those technology supports with what is used in the classroom.

Pisacreta (2013) conducted a small scale usability study in which students ages 11-13 were asked to complete two typing tasks on a tablet—one using an onscreen keyboard and one using an external keyboard. In addition they were asked to complete a small set of items that required revision and text editing with either the onscreen keyboard or the external keyboard. They found that typing paragraphs using the on the onscreen keyboard was slower and less accurate for most students. They also found that students had difficulty highlighting, copying or pasting text on the tablet (actions that did not require the use of a keyboard). Strain-Seymour, Craft, Davis, & Elbom (2013) found similar results in another small scale usability study, but noted a difference in preference and facility for the onscreen keyboard for younger students and those students who had not yet had formal keyboarding training.

Davis, Orr, Kong, & Lin (2015) conducted the first large scale study of writing assessment with a sample 5th grade and high school students. Each student was provided with either a laptop, a tablet, or a tablet with an external keyboard and asked to respond to a grade level appropriate essay prompt. Results indicated no difference in the essay score or surface level essay features across study conditions. In many ways this is counter to the expectations that most adults might have given their own experiences in using onscreen keyboards. What is even more surprising is that these findings seem to hold across both grade levels studied. The Strain-Seymour, Craft, Davis, and Elbom (2013) research would have suggested that older students who had more training and experience with keyboarding skills might have been expected to struggle more with the onscreen keyboard than younger students who had not yet developed this facility. While the survey responses of the high school students did indicate a definite preference for a physical keyboard, this preference did not translate into a performance difference across conditions. In addition, while high school students' perceptions of the virtual keyboard were not as strongly positive as their perceptions of the physical keyboards, neither were they completely negative as 71% of high school students reported finding the onscreen keyboard somewhat or very easy to use.

Science

Device effects in science have been less studied than other subject areas. Davis, Kong, & Orr (2015) conducted one of the only studies to have psychometrically evaluated science assessment content when delivered on computer and tablets. They found no evidence for statistically significant differences between tablets and computers. Yu, Lorié, & Sewall (2014) evaluated both biology and algebra items with high school students through a cognitive lab and found stronger associations with item type than they did with content area for measures of student certainty and frustration. Science holds one of the greatest potentials across all subject areas for leveraging technology to enhance measurement. It will be important to continue to monitor device effects in this area as new item types such as simulations are introduced into assessment.

Research by Item Type

Multiple Choice Items

Research on device comparability for multiple choice items shows them to be very robust to use on different devices with little to no observations of device effect (Davis, Strain-Seymour, & Gay, 2013; Olsen, 2014; Davis, Kong, & McBride, 2015). Davis, Strain-Seymour, & Gay (2013) noted some need for students to adapt their approach from a paper to a digital display of a multiple-choice item as some students did experience slight precision issues when using their finger on a touch-screen display if they tried to precisely select within the answer option "bubble." They did not realize that touching anywhere on the text of the response option would select the answer and, as a result, they had slight difficulty pinpointing the "bubble" to make their selection. Because the roll-over effect of the cursor switching to a pointing finger (indicating that a wider area could be selected) did not appear on the tablet, students did not benefit from this additional information. Strain-Seymour and Davis (2013) also found that few students (either those who tested on tablet or computer) recognized that you can select an answer choice by clicking anywhere on that choice. However, these issues are easily addressed through practice with tutorials and do not appear to cause much frustration for students.

It should be noted that the multiple choice response format itself can vary with the response options being either simple text (more typical) or complex interactions which include art, graphs, charts, etc. With these more complex interactions attention should be paid to the visibility and clarity of the response options themselves. Large images which introduce scrolling or make it difficult to view the item stem at the same time as the response options may add difficulty for students when using devices with smaller screen sizes (such as tablets).

Multiple Select Items

Multiple select items generally perform similarly to multiple choice items across devices. Strain-Seymour and Davis (2013) found that students did not recognize the user interface convention of using checkboxes instead of radio buttons to indicate that multiple answers can be selected. Students consistently used cues from the item, such as "choose two" with the word "two" bolded. Thus, language within items should be consistent in indicating how many answer choices should be chosen or that multiple can be chosen (e.g., "Choose all that apply."). Davis, Kong, & McBride (2015) found no evidence for device effects for multiple select items for either reading or mathematics items.

Drag and Drop

Drag and Drop as an item type covers a wide array of item behaviors. Items of this type are typically defined by having "draggers" and "bays." Draggers are objects that students manipulate to create their response. Bays are the fixed locations where students drop the draggers. However, there can be different styles of drag and drop items that involve matching a single dragger to a single bay or which allow for reuse of draggers to drop into multiple bays. Draggers can be text objects or art objects. The scoring of drag and drop items can enforce certain orderings or sequences for credit or allow credit for any response which includes the correct set of draggers in any order.

Davis, Strain-Seymour, and Gay (2013) found that students interacted well with draggable objects on the tablet, with some commenting favorably on the ability to interact directly using their finger and not having to use a mouse as an intermediary to express their intent. However, when the "target area" for dropping the object was small or close to other target areas, students sometimes struggled to precisely place the object (Davis, Strain-Seymour, & Gay, 2013; Strain-Seymour, Craft, Davis, & Elbom, 2013). As a result, the objects might "snap back" to their original position rather than sticking to the place where the student intended to drop the object. This snapping effect sometimes caused students to question whether their answer was correct; some thought that the tablet was giving them a hint that their answer was wrong. Conversely, when the target area was sufficiently large or separated from other target areas the question was relatively easy for students to interact with.

While the action of dragging an item is inherently natural on a tablet, the specific design of the user interface for drag-and-drop items plays a significant role in students' understanding of the item type. Strain-Seymour and Davis (2013) identified several usability issues for students with drag-and-drop items. The first issue was recognizing the item as a drag-and-drop vs. a hot spot or text entry response. The drag-and-drop items had drop-zones that expanded to fit draggers as they were dropped on the drop-zones. For some students, it initially appeared as if the draggers would not fit into the drop-zones, which may have further impeded recognition of the item as drag-and-drop through non-textual cues. While drop-zone re-sizing to accommodate draggers does help to preserve space and avoid cuing around how many draggers could be assigned to a drop-zone, students may be confused when the dragger is significantly larger than the drop-zone. In addition to it looking like the dragger will not fit, the fact that the dragger may entirely cover not just one drop-zone but all drop-zones makes it very difficult to drop the dragger in the right location. Most students were able to recognize the item as drag-and-drop and recognize draggers as draggers and drop-zones as drop-zones — either immediately or eventually — by careful reading of the item rather than by visual appearance of the draggers and drop-zones.

Additionally, drop zones may be best placed at the top with draggers below rather than the reverse. In the Strain-Seymour and Davis (2013) study the drop-zones were at the bottom of the screen at the point where the scrolling started added a further complication. Working with a drag-and-drop item within a scrolling screen will typically present usability problems as all elements are not visible at the same time and have to be dragged across the "fold." This issue and the proximity of the drop-zones to one another were what caused the most frustration amongst high school students. When positioning a dragger over a drop-zone, some type of user feedback should be provided, such as the drop-zone changing in appearance, to show what drop-zone a dragger will snap to if dropped while in its current position. While careful attention should be paid to the interface design for drag and drop items, this does appear to pay off in greater device comparability. Davis, Kong, & McBride (2015) using an interface revised based on earlier research found no evidence for device effects for drag and drop items.

Hot Spot

In some instances, a hot spot item functions as a technology enhanced version of a multiple choice or multiple select item with students selecting responses by highlighting an object rather than selecting a radio button. Some hot spots may involve more complex interactions though selecting "hot" areas of a chart or figure. Davis, Strain-Seymour, and Gay (2013) found that most students experienced no difficulty in placing a mark or selecting an image using their finger within a hot-spot question. Strain-Seymour and Davis (2013) found that the hot spot should switch to its selected state on clicking or tapping down rather than on release in order to appear more responsive, reduce confusion over the hot spot being draggable, and avoid double-clicking or long-click behaviors on the tablet, which can cause additional functionality (zoom, text selection magnifier) to be triggered accidentally. Davis, Kong, & McBride (2015) found no evidence for device effects for hot spot items.

Inline Choice

Inline choice items are also sometimes referred to as "cloze" items or "drop-down" items as they allow a student to respond to an item by selecting a response from a drop-down menu which may be embedded inside of a sentence or paragraph. Strain-Seymour and Davis (2013) found that inline choice did not present issues to students in a small scale cognitive lab/usability study. This was confirmed by Davis, Kong, & McBride (2015) in a large scale evaluation of device comparability.

Fill-in-the-Blank (Short Response)

Fill-in-the-blank items are similar to inline choice, but require text entry for students to respond rather than selection from a pre-specified list. Some assessment programs have adapted this format by creating a "fill-in" box that only allows a limited range of characters to be used (e.g. only numeric responses). Others allow a full range of alpha-numeric and even special character symbols to be used. As text entry is required to provide a response to this item type the type of keyboard students are using becomes an important factor. In the case of numeric only responses Davis, Strain-Seymour, and Gay (2013) found that the onscreen keyboard native to the tablet defaulted to the alphabet even though the answer allowed only numbers. Students had to know or figure out how to navigate to the numeric/symbol keyboard to enter their answer. In the case of alphabetic responses, students adapted quickly to text entry with the onscreen keyboard—typically used single or double-finger hunt-and-peck style typing to enter their responses. Most students did not appear to be excessively bothered by this and were able to complete the question with this approach. Many students remarked that they liked the onscreen keyboard though a few students commented that this would be more of a concern if they had to type more than a few words. However, students did seem temporarily dismayed or inconvenienced by the fact that the opening of the onscreen keyboard pushed question content off-screen. Students had to scroll back up to enter their response or to see information they needed to answer the question.

Strain-Seymour and Davis (2013) found that students did not seem to have issues typing numbers into a blank using the on-screen keyboard on the tablet. Some younger students without tablet familiarity may have taken a moment longer to locate the number keys. Students either tapped somewhere on the item or used the keyboard-dismiss button to put the keyboard away before navigating to the next item. Davis, Kong, & McBride (2015) found no evidence for device effects for fill in the blank items.

Graphing

Graphing items are those that require students to plot data or create graphs of mathematical functions using specific graphing interfaces and tools. Some items may include graphs but require students to interact with the graphs through other mechanisms (e.g. drag and drop, hot spot, etc.) and are not classified as "graphing items." Graphing items may take many forms including line graphs, bar graphs, scatter plots, function graphs, etc.

Within graphing items the key factor is for the student to have good visibility of the graph while creating it using either the mouse or the finger on a touch-screen device. This includes visibility of the graph structure itself (e.g. Cartesian coordinates; Strain-Seymour & Davis, 2013) so that dots, lines, or bars may be precisely placed in the correct and intended location with confidence. For example, Davis, Strain-Seymour, and Gay (2013) found that bar graph questions worked well on tablets so long as the width of the bar was large enough for students to drag with their finger and the bar was sufficiently separated from other bars. Thinner bars or bars that were spaced too close together proved challenging for students as they were unable to grab the bar at all or inadvertently grabbed the wrong bar. Similarly, they found that point or line graph questions requiring students to plot points and lines on a Cartesian coordinate graph were sometimes challenging for students because the student's finger obscured the placement of the point. As a result, some students expended additional effort to move their points to the correct locations. Holz and Baudisch (2010) similarly observed that touch-screen input accuracy may suffer when the finger blocks some part of the graphical interface.

In a study with students in grades 4, 8, and 10, Strain-Seymour and Davis (2013) evaluated an interface change made to better accommodate graphing items on the tablet called the "halo." A halo is of a semi-transparent circle around a point, which can be used to move the point around without having one's finger block the view of the point's position in relationship to the lines below (see Figure 6). They found that the halo worked for tablet users and did not it create a distraction or a source of confusion for computer users.

Tablet users could see the halo extending out beneath their finger and, thus, were not observed lifting their fingers to check if the point was still there. Another issue seen previously involved a student sometimes not succeeding in picking up the point while attempting to drag. The student would pick up his or her finger only to realize that the point was no longer there. Students found it easier to drag with precision when using the halo as a guide.

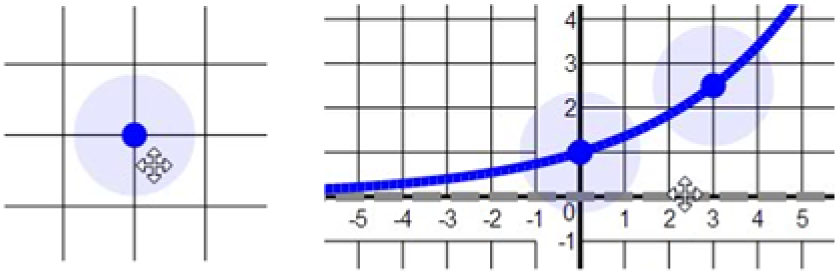
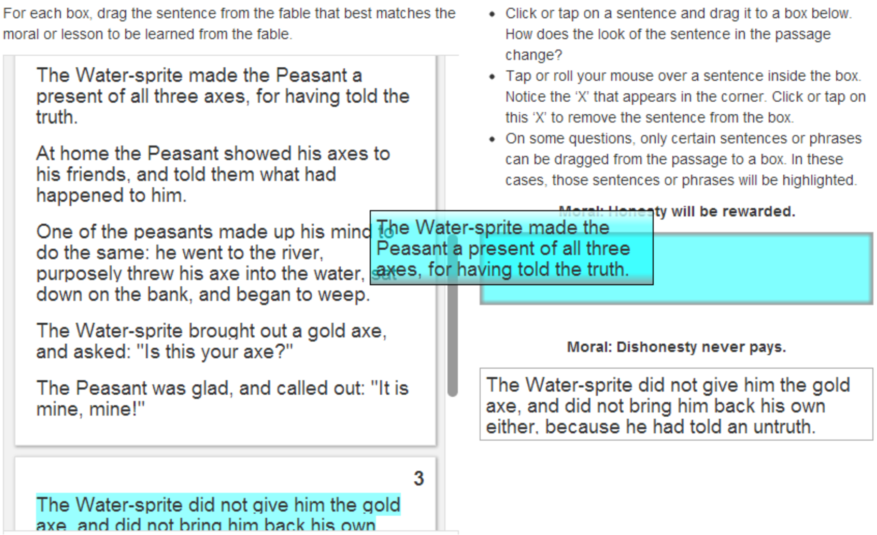


Figure 6: Example of a halo as used in graphing items

Text Extraction

Text Extraction allows drag-and-drop type response interactions to be applied to a reading passage (see Figure 7). A passage may have extractable selections of text. These chunks may consist of all sentences, all paragraphs, etc. Alternatively, only certain words, phrases, or sentences might be eligible for extraction. To create a response, a test-taker pulls a text piece out of the passage and drops it into a labeled box. Strain-Seymour and Davis (2013) found that students did not always recognize a text extraction item as such, even after encountering it in a tutorial. They suggested that the boxes for the extracted text might be a place to focus, since these boxes were initially, but only briefly, mistaken by a couple of students as boxes to type text into in the fashion of an open response item.



For each box, drag the sentence from the fable that best matches the moral or lesson to be learned from the fable.

The Water-sprite made the Peasant a present of all three axes, for having told the truth.

At home the Peasant showed his axes to his friends, and told them what had happened to him.

One of the peasants made up his mind to do the same: he went to the river, purposely threw his axe into the water, sat down on the bank, and began to weep.

The Water-sprite brought out a gold axe, and asked: "Is this your axe?"

The Peasant was glad, and called out: "It is mine, mine!"

3

The Water-sprite did not give him the gold axe, and did not bring him back his own

- Click or tap on a sentence and drag it to a box below. How does the look of the sentence in the passage change?
- Tap or roll your mouse over a sentence inside the box. Notice the 'X' that appears in the corner. Click or tap on this 'X' to remove the sentence from the box.
- On some questions, only certain sentences or phrases can be dragged from the passage to a box. In these cases, those sentences or phrases will be highlighted.

Moral: Honesty will be rewarded.

The Water-sprite made the Peasant a present of all three axes, for having told the truth.

Moral: Dishonesty never pays.

The Water-sprite did not give him the gold axe, and did not bring him back his own either, because he had told an untruth.

Figure 7: Example of a Text Extraction Item

Text Highlighting

Text highlighting allows for student responses to be indicated by effectively making all or portions of a reading passage “hot spots” that can be selected and highlighted. A passage may have highlightable sentences, paragraphs, poem lines, or stanzas. A single click or tap highlights the text. A click or tap to an already highlighted text chunk un-highlights it. The predetermined chunks available for highlighting facilitate machine scoring, as opposed to free-form highlighting. Strain-Seymour and Davis (2013) found that all students taking the test on computer immediately recognized how to highlight sentences due to the roll-over effect of the cursor. Most but not all tablet users recognized how to highlight. They recommended adding text to the item directions that explicitly gave direction on how to highlight.

Equation Editor

Equation editor items are those that allow for special math symbols for a constructed response of math/numeracy content either alone or in conjunction with explanatory text. Equation editors have been commonly available in commercial software products for some time, but were designed to be used with a mouse and keyboard. Yu, Lorié, & Sewall (2014) recommend that changes be made to these interfaces before implementing on tablets. In their research they found that the equation editor was confusing to many students when used on tablets. Specifically students were concerned about positioning the cursor within the equation editor interface and usability of buttons to create superscripts and subscripts. As with all touch-screen interfaces, the size and spacing of buttons should be carefully evaluated to make sure students can precisely and accurately select the intended function.

Device Positioning and Ergonomics

Tablets are inherently flexible devices intended to support a range of different possible positions for use. Research with adults (Young, Trudeau, Odell, Marinelli, Dennerlein, 2012) indicated that tablet users take advantage of their devices' many potential display positions, changing device position and their bodily position based on the task. However, in a study by Davis, Strain-Seymour, and Gay (2013) the majority of students were observed to place the tablets flat on the table and leaned over them to view the screen despite being given no specific direction about how to position the tablets. A few students held the device throughout the test (more likely with the smaller 7-inch tablet used in the study than with the larger 10-inch tablet), and others propped the tablet up on its cover. Yu, Lorié, & Sewall (2014) reported that students in their study were split between laying the tablet flat on the table and propping it up at an angle. They further found a relationship between self-reported level of onscreen typing skills and tablet position with those students who reported having "Advanced" onscreen typing skills preferring to keep it flat on the table.

However, Davis & Strain-Seymour (2013a) observed greater variety in how students positioned the tablets (see Figure 8). While laying the tablet flat on the table was still the most common position, some students propped the tablet up on its case and used it at an angle, while others propped the tablet up on one arm while using the other hand to navigate and answer questions. Yet other students sat back in the chair and held the tablet in their laps. Students were given no specific instruction about how to position the tablets, and when asked about their choice of positions, most students reported that this is how they use a tablet at home. Differences in the findings of the studies are likely due to the study settings: school classroom in the first two cases vs. conference room in the latter case. This difference may be responsible for the increased observation of flexible positioning as, within the school setting, students may have expectations about appropriate seating/materials positioning. Additionally, chairs in the conference rooms were cushioned and swiveled, which allowed greater flexibility to lean back and might have encouraged different behaviors than what would be seen in a classroom.

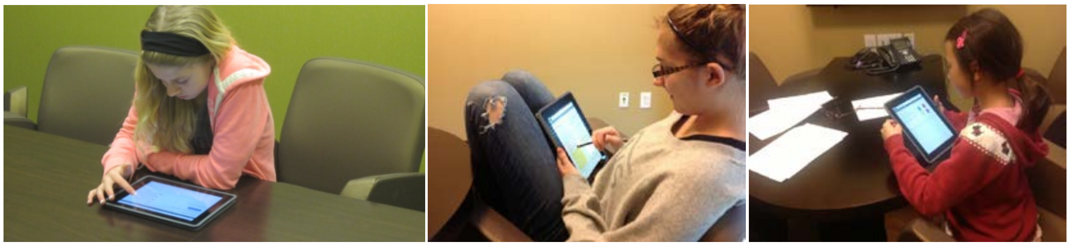


Figure 8. Different device positions for using tablet to access test content

Impact of External Keyboard on Device Positioning

A decision to require the use of an external keyboard has direct impacts on the different device positions available to students. Strain-Seymour, Craft, Davis, & Elbom (2013) found that while students appeared to be comfortable working with the tablet flat on the table when using the onscreen keyboard, the addition of the external keyboard caused observable awkwardness, generating several student comments. Once the external keyboard was added, some students lifted the tablet at an angle, and then set it back down, as if looking for a way to prop it up. With the external keyboard, greater degrees of head and eye movements were observed, as most students looked at their fingers during typing and then up at the essay text box, scanning back and forth as they worked on their essays. As a result of these findings, Davis, Orr, Kong, and Lin (2015) provided a tablet stand for every student in the external keyboard condition of their study (see Figure 9).

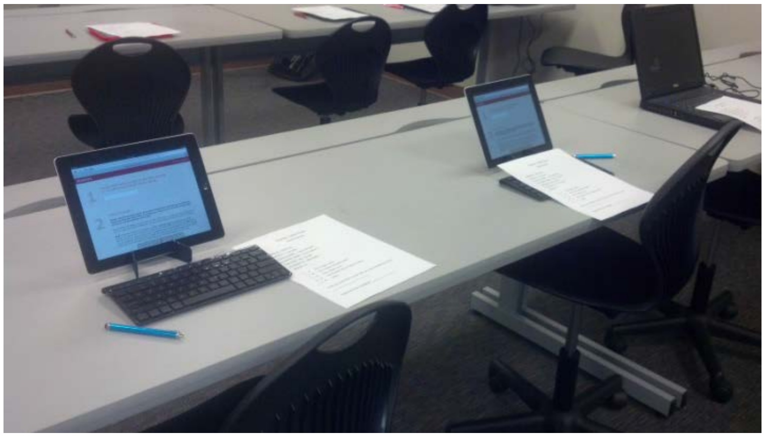


Figure 9. Tablet position in stand when paired with external keyboard

Ergonomics: Screen glare, Eye fatigue, Neck and Back Strain

Davis, Stain-Seymour, and Gay (2013) evaluated ergonomic impacts of device positioning. While most students in their study were able to complete the study items without significant strain or difficulty, the short duration of the testing session (30-45 minutes) may have limited the degree of discomfort students experienced. Some students mentioned that they would likely suffer some issue such as neck pain, thumb strain, or headache due to holding or viewing the device for a lengthy testing session. A few students also mentioned that the angle of the device in conjunction with overhead lighting or eye glasses may create a glare that would cause eye strain.

Switching between device positions

Encouraging flexibility in shifting tablets from one position to another during testing may help alleviate some of these ergonomic issues (Davis & Strain-Seymour, 2013a). If students have the opportunity to work with tablets as part of daily classroom activities, they'll likely have a greater familiarity with the device and the ways its position can be changed for particular tasks. However, the more students are allowed to move the tablet to different positions, the more likely it is that other students in the testing room might be able to view their responses either intentionally or unintentionally. Additionally, depending upon the type of keyboard, case, and stand selected, it may be more or less difficult to change positions. For example, if a folio type external keyboard is selected students would have to remove the tablet from the case entirely in order to allow for hand-held use.

Use of Other Peripherals

Stylus and Tablet Mouse

Styluses are small, reasonably affordable, resemble a pen or pencil, and while they are not technically more precise than a finger, they allow somewhat better visibility since the students' hand is not blocking the screen (Pogue, 2012). In general very few students have experience using a stylus with a touch-screen device (Davis, Strain-Seymour, & Gay 2013; Yu, Lorié, & Sewall, 2014). Strain-Seymour and Davis (2013) found that it was particularly tempting for stylus users to treat the stylus like a highlighting pen and pull it across the text but when this did occur, students recognized that highlighting occurred with just a tap. Yu, Lorié, & Sewall (2014) included a stylus in their study of student responses in Algebra I and Biology assessments to see if it would assist with graphing precision. They found that among students who had not previously used a stylus there was a preference to use their hands and fingers, as it was easier and more natural. Some students thought using the stylus would make it more difficult or more complicated to interact with the touch-screen. Davis, Orr, Kong, & Lin (2015) evaluated the utility of a stylus for text editing and revision in written composition items. The initial hypothesis was that students would use the stylus to assist them in more precisely placing the cursor (compared to just using their finger) as they went back to revise their compositions. Actual student use of the styluses differed from this expectation in that most students did not have experience using a stylus either at home or in school. As such this specific use for the stylus did not appear to occur to them. Of high school students who were provided a stylus, 39% did not use the stylus at all and 22% reported that they did not find it helpful. Fifth grade students reported a more favorable use and impression of the stylus, but researcher observation was that they primarily used it as an aid to strike the keys on the onscreen keyboard and did not use it to assist with cursor placement as part of a revision activity.

Conclusions from these research studies are consistent in recommending that use of a stylus be optional for students so that it is available for those who are familiar with it and want to use it but does not interfere with the ability of other students to interact with the device. Davis, Strain-Seymour, & Gay (2013) go so far as to suggest that it is not necessary to require a stylus for student testing for selected response item types as long as specifications for content development include best practice for touch screen interfaces (minimum object size, spacing, etc.). In any case, availability of a stylus is likely to be attractive to some students and so offering a stylus as a peripheral should be considered. However, students (especially younger students where fine motor control is more of an issue) should have an opportunity to work with a stylus outside of testing as part of their regular academic instructional use with tablets.

A tablet mouse may be connected to a tablet via USB cable or Bluetooth and incorporate tablet specific features such as use on almost any surface for portability and multi-direction scrolling (see Microsoft's wedge mouse or Apple's magic mouse; Brown, 2012; Stern, 2010). There is, to date, no research on use of tablet mice as peripherals for assessment owing in large part to their lack of general popularity as tablet peripherals.

Tablet Stands and Cases

In selecting tablet stands and cases, it is important to consider the position in which students will use the device as well as the need or desire to change device positions throughout the assessment period. As suggested, earlier most students feel comfortable using the tablet in a flat position for drag, swipe, and tap intensive work. Therefore, the position that is most critical for a tablet stand to support is the upright position for use with an external keyboard (Strain-Seymour & Brock, 2014). While some tablets have cases which can be folded to prop up the tablet, many students struggle with the folding covers, find them to be insufficient for longer periods of use, and are not able to gain the degree of vertical support desired for using with an external keyboard. The easel type stands (see second image in Figure 10) can be a good choice, as they also tend to be inexpensive and resilient. Choices may be different with a 1:1 initiative, since daily usage will encourage familiarity with even more complex tablet stands design. It is also important to keep in mind that if a tablet is already in a case which supports its use without an external keyboard, it may need to be removed from that case to fit into an external stand. Lastly, a stand should have good support from behind the upper portion of the screen. Even as students are using the keyboard, they may still be tapping on portions of the screen when editing.

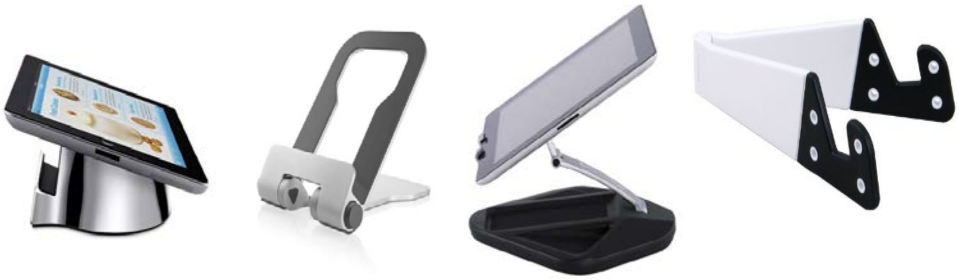


Figure 10: Examples of different tablet stands

Device Set-up and Security

Tablet Configuration for Testing

As with all digital devices, tablets must be configured for testing purposes prior to the assessment administration. This includes downloads of required software applications, configuration of settings for onscreen keyboards, charging of devices, and preparation of tablet peripherals. This setup can be challenging to accomplish if tablets that are issued to students in 1:1 programs are also used for testing purposes as the school technology staff may not have physical access to these devices prior to testing day. Additionally, tablets which are issued to students may have unexpected damage (screen cracks, battery issues, etc.) that could impact their use for testing. Many software applications used for testing will require the tablets to be put into "single application mode" so that students cannot access other applications either locally or through the internet during test administration. Schools should plan in advance for preparation of tablet devices and should have spare tablets available on testing day should issues arise on testing day (Strain-Seymour & Brock, 2014).

If external keyboards are required for tablets, it is important to know whether they connect wirelessly (e.g. through Bluetooth) or physically to the tablets. External keyboards which require wireless pairing must be physically co-located with the specific tablet to which it has been paired. It is incredibly easy for keyboards to get swapped amongst students or classrooms and thereby require extra effort to re-pair.

Testing Room Set-up

While hardware and software configuration is important, the physical set-up of the testing room itself must also be considered. Students must have sufficient working space to enable use of digital as well as physical resources. As it turns out, screen real estate is not the only real estate to be concerned about. While the best case scenario in terms of physical space is one in which nearly everything students will need for the assessment is on the tablet, this is often unrealistic from a usability perspective where students may need to reference multiple items at one time (Strain-Seymour & Brock, 2014). As a result there is still a potential need to allow for physical space for the tablet itself, a tablet stand, an external keyboard, scratch paper and pencil, a physical calculator, etc. Making sure that students have sufficient "desk" space to spread these materials out is critical.

Additionally, testing rooms must be set-up to promote security during the testing administration. This might mean allowing sufficient space between students to preclude copying, use of card-board dividers (see Figure 11), etc. The types of physical environments in which students might use devices for instructional activities (working collaboratively or in small groups) are not likely to be appropriate for use during test administrations.

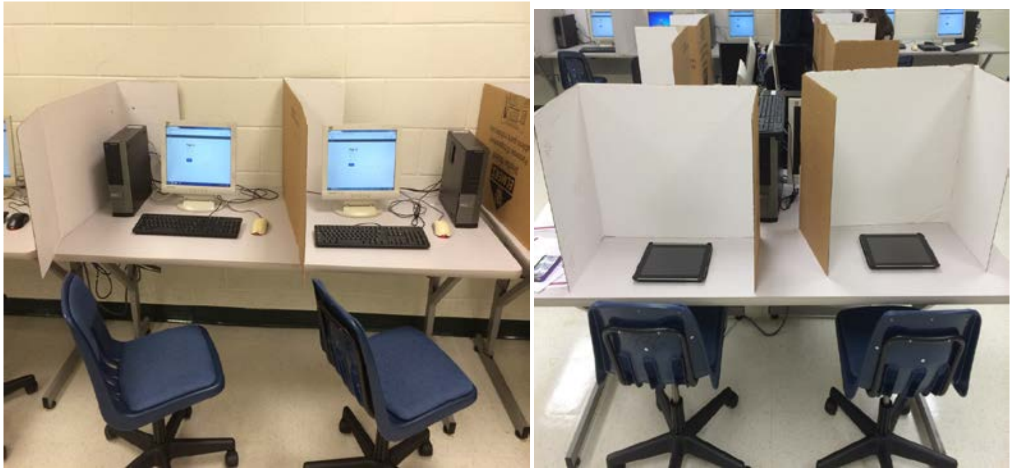


Figure 11. Use of cardboard dividers to create testing stations for desktops and tablets

Differential Effects by Student Subgroups

Given that touch-screen tablets in the modern sense have only been available since the launch of the iPad in 2010, research into the area of device comparability for student subgroups using tablets is limited. As such, the focus of assessment research to date has largely been on overall group device effects. There has been some research to support development of assessment applications specifically for special populations which make use of tablet technology (see for example Pearson's Tests of English Language Learning (TELL) assessment; Hildago & Sato, 2014). However, the relationship between student group membership and score comparability across devices is a very important, but, as of yet, unexplored domain.

Differential Effects Based on Device Familiarity and Experience

Academic use of digital devices is increasing as 1:1 technology initiatives (1 device for each student) and concepts such as the flipped classroom become more commonplace in educational pedagogy (Hamdon, McKnight, McKnight, & Arfstrom 2013; McCrea, 2011). However, surveys of students conducted across various research studies suggest that more students use touch-screen devices at home than in school. For example, Davis, Orr, Kong, & Lin (2015) found that a large percentage of students in their study reported using both computers and touch-screen tablets regularly at home (48% of 5th grade and 64% of high school students). However a majority of students reported using a computer only in school (59% of 5th grade and 51% of high school students). Similarly, Davis, Kong, & Mcbride (2015) found that more students reported using tablets and smart phones for personal use than for school work (44% vs. 28% for tablets and 84% vs. 31% for smart phones) though rates of usage for desktop and laptop computers were more similar between personal use and school work (27% vs. 36% for desktop computers and 61% vs. 66% for laptop computers).

Interestingly, Strain-Seymour and Davis (2013) found that the students who had tablet experience generally had the same types of usability issues as students who did not. However, students with tablet experience did differ from other students in their recovery from or non-reaction to the appearance of unintended functionality. For instance, when an accidental double tap led to the screen zooming in, the students with tablet familiarity quickly adjusted it back to 100%, while students without tablet experience took slightly longer to recover and re-adjust the screen. Similarly, when a long tap led to the appearance of the text selector magnifier, tablet users recognized it and were unbothered by it. Other students generally did not react poorly to the text selector, but might have noted its appearance without knowing what it was before proceeding to ignore it.

In sum, research to date suggests that tablet familiarity is a critical factor for students in using touch-screen devices successfully for assessment (e.g., use of pinch-and-zoom to compensate for reduced input precision, locating numbers on keyboard, etc). However, while familiarity with touch-screen devices is important, the amount and type of exposure required for success may be less than what might be expected. In other words, general familiarity with touch-screen functionality combined with use of assessment tutorials to familiarize students with specific item types and testing software conventions may be sufficient to overcome any potential concerns. Additionally, careful attention to the design of the user interface for the testing application is important for all students regardless of their familiarity with touch-screen devices. For example, Strain-Seymour and Davis (2013) found that despite the familiarity of students in the study with drag-and-drop items, these were typically the least usable items in the study because of the way in which the user interface was designed.

Differential Effects by Age or Grade Level

Many of the research studies of which deal with touch screen user interface design, in general, have been conducted outside of educational contexts and are frequently conducted with adult software users (see for example, Findlater & Wobrack, 2012; Young et al., 2012). Good design principles established with adults typically translate well into usable designs for students. Within the research studies conducted with school-aged children in the educational assessment context, however, two primary findings have emerged relative to student age and grade level. The first is in relationship to student facility with and preference for onscreen or external keyboards for tablets.

The second is with regard to how students respond when they encounter a usability issue with the device or user interface of the testing software

In the first case, multiple studies have identified a stronger preference by younger students (typically 4th and 5th grade level) for the onscreen keyboard while older students (typically middle and high school students) tend to prefer the external keyboard (Davis, Orr, Kong, & Lin, 2015; Davis, Strain-Seymour, & Gay, 2013; Strain-Seymour, Craft, Davis, & Elbom, 2013). These preferences are hypothesized to be related to the strength of students' keyboarding skills with stronger keyboarding skills resulting in lower preferences for the onscreen keyboard. Younger students were frequently observed to have less proficient keyboarding skills, while older students tended to have more proficient keyboarding skills. Strain-Seymour, Craft, Davis, and Elbom (2013) report that younger students who might use a "hunt and peck" method to select keys with one or two fingers were either equal in speed on both keyboards or slightly faster on the on-screen keyboard. Keyboarding input for older students was markedly slower on the on-screen keyboard, with the differential between the two speeds being greater the better a student's typing skills were. Davis, Orr, Kong, and Lin (2015) confirm the differences in keyboard preference between younger and older students, but importantly note that these preferences did not translate into differences in performance on written essay responses.

In the second case, Strain-Seymour and Davis (2013) found that younger students were slightly more likely to encounter usability problems than older students and that younger students seemed to recover less easily and were more likely than older students to skip the item after encountering a usability issue . They reported that older students were more willing to persist, try the same thing again, or try a different approach, when faced with a usability issue. These observations were made both for students testing on computers and tablets—though slightly more usability issues were identified for tablets in this study. Similarly, Davis (2013) found that elementary aged students were more likely than middle and high school aged students to report that they were uncertain how to navigate from one item to the next and to ask for help in doing so. By contrast, high school students were more willing to experiment with the different icons and buttons in the software and usually figured out how to navigate to the next item without prompting from the study facilitators. It should be noted that in both studies, students had not practiced with the testing software in advance of participating in the research. It is likely that exposure to the software controls and device functionality through tutorials or practice assessments would address most of these concerns.

Conclusions

The increasingly popularity of touch-screen devices like tablets and smartphones creates new opportunities to engage students with digital content in the classroom and for assessment which will be more closely aligned with the way they experience the rest of the world. Although these opportunities are very exciting and open up new potential methods for measuring constructs of interest, the comparability of scores and of student experiences across devices is an important consideration for online assessment programs. Projecting forward in time, the number of different types, sizes, makes and models of digital devices in use within schools is likely to continue to expand as new technology offerings blend with devices previously purchased. It is important to have a framework to sort device features that are likely to have substantive impacts to student experience and assessment performance from those that may only have minor impacts or no impacts at all. The concept of device "form factor" as discussed in this paper is critical to helping to identify areas where comparability research may be needed as technology continues to evolve.

While the use of touch-screen devices for assessment introduces a series of new and important considerations for test developers, the research to date suggests that assessment content can be delivered comparably across devices with appropriate attention to user interface and technology requirements. Certainly, there are no studies which have identified large statistically or practically significant device effects. Any effects observed have tended to be small and isolated to specific item types or interactions. Additionally, students themselves appear fairly robust to device effects and adapt quickly to new methods of interaction. In considering different subject areas, reading and writing should be given careful consideration so as to structure the user interface and device requirements to provide the optimal experience for reading texts and creating written text. There is great opportunity with mathematics and science to incorporate technology enhanced item types that allow for deeper levels of cognitive complexity to be assessed as students are asked to "categorize", "connect", "create", "order", "match", "plot", "rearrange", and "transform". However, individual item interactions need to be carefully reviewed on each type of device that will be allowed for testing so that any potential usability issues or areas of confusion for students can be identified.

The research summarized in this paper should not be interpreted to suggest that tablets should only be used as a means of expanding the set of available devices on testing day. Student familiarity with tablets in an academic context is crucial and tablets are best used as part of a technology rich learning environment throughout the school year. In addition, it is important for students to have access to tutorials and practice tests that allow them to become comfortable with the specific user interface of the testing software and the allowable and expected interactions of the specific item types used in the assessment. Teachers should also have access to tutorials and practice tests and, to the extent feasible, be encouraged to create classroom quizzes and tests which mirror the functionality of the summative assessment. Professional development for the teaching staff is an important component to an online assessment program as the adults in the system are often more intimidated by the technology than their students.

Recommendations for ACARA

As ACARA, moves to online assessment with the inclusion of tablet devices, there are several key factors which should be considered.

1) Use of External vs. Onscreen Keyboards

   While many assessment programs are requiring the use of external keyboards for tablets, this requirement alone does not ensure a good user experience for students and may negatively impact students in some circumstances. In making this decision, ACARA should consider the amount of text generation that students will be asked to produce for each item type. For short (one word or single phrase) responses, an external keyboard does not appear to be necessary. For longer (extended) responses, this might be considered as an option which schools make available to students but would function like scratch paper—available if the student wants it, but not required. Any requirement to use an external keyboard should specify technical specifications for the type and size of keyboard as well as accompanying tablet stand and should also take into consideration whether the external keyboard would be required for all sections of the assessment or only for certain sections. Lastly, schools should be required to provide evidence that the students had an opportunity to use the external keyboard as part of regular classroom activities and that they weren't provided only for the test sessions.

As part of their research efforts, ACARA should be encouraged to take note of the types of external keyboards students use and whether there is any relationship between certain keyboard types and student success or challenge in using them. Specific attention should be given to younger students versus older students with the potential for different policies for different age groups. Additionally, study facilitators (invigilators) should look for instances where students are revising their written responses, observe whether there are challenges with this process as well as how students manage the revision process. Lastly they should probe with students during the structured interviews around whether they would have revised their responses if they had different tools available (e.g. a laptop with a mouse).

Within study conditions which make use of onscreen keyboards study facilitators (invigilators) should note how easily students are able to navigate between alpha- and numeric- keyboards, how easily they are able to open and close the onscreen keyboards, and the different typing techniques that students attempt. They should probe with students during the structured interviews relative to their experience with typing as well as any use of auto-correct, autocomplete, split screen keyboarding, etc. (if those configurations will be possible during the test administration).

2) User Interface for Drag and Drop Item Types

Given the diversity of different drag and drop interactions and the previously observed usability challenges with this item type, ACARA is encouraged to pay specific attention to this item type in its research studies. Study facilitators (invigilators) should be made aware of the different types of drag and drop interactions and encouraged to probe with students during the structured interviews for elements of their interactions which were particularly challenging.

3) Scrolling for Reading Passages

While ACARA is making efforts to avoid scrolling within its reading passages, some of the longer passages at higher levels will have scrolling. Study facilitators (invigilators) should observe whether students can identify when a passage scrolls and the ease with which they scroll through the passage. The specific mechanism for scrolling (use of a finger within the passage vs. scroll bar, etc.) should be noted if there are multiple options for this interaction. During the structured interviews, study facilitators (invigilators) should probe with students to understand their perception of the scrolling activity and ask whether students understood how to scroll if needed.

4) Tools

From the sample tests ACARA provided, it is unclear whether students will have access to non item specific tools. If they will, the usability of these tools for tablet delivery should be evaluated. Study facilitators (invigilators) should observe student use of these tools (both frequency and ease of use) and probe with students during the structured interviews for any areas of specific challenge. If students will not have access to tools within the testing interface, study facilitators should probe with students to find out if specific tools would have been helpful and whether not having those tools created challenges for them. While an absence of online tools would be a consistent factor across devices, this absence may be more keenly felt for users of one device more than the other. If students have physical tools (scratch paper and pencil, physical calculators, etc.) study facilitators (invigilators) should probe to find out how easy it was for students to transfer information from the physical tools to the online interface.

5) Device Positioning and Ergonomics

Study facilitators (invigilators) should make note of the different positions in which students use the tablet devices as well as whether they change positions throughout the testing session. They should probe with students during the structured interview to identify any ergonomic impacts related to physical strain or eye strain. If students are using tablets in a vertical position within a tablet stand, study facilitators should inquire how easy it was for students to interact with the touch screen from that position.

6) Touch Screen Specific Features

If features such as pinch-and-zoom and screen rotation are enabled study facilitators (invigilators) should note how often they are used. Based on previous research it seems unlikely students will rotate the tablet, but pinch-and-zoom to enlarge certain details may be observed. It will be important to note under which circumstances students used pinch-and-zoom capabilities (e.g. in reading passages; for reviewing charts, etc.) and whether they found those capabilities helpful in viewing objects within the test.

7) Audio Functionality

No research to date has looked at the comparability of audio functionality across devices. Given that ACARA will be utilizing audio prompts to measure spelling, specific attention should be given to student interactions with these prompts. Study facilitators (invigilators) should note whether students have any difficulty in playing or replaying the audio from their specific device, whether there are volume control issues which may be present for one device or another, and whether students tend to replay (if allowed) more frequently on one device than another.

References

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

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments (APA) (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.

Ash, K. (2013). Reading in the age of devices. *Education Week, 6(2).* Pages 22-23. Retrieved from: http://www.edweek.org/dd/articles/2013/02/06/02reading.h06.html

Ballagas, R., Rohs, M., Sheridan, J. & Borchers, J. (2004). BYOD: Bring Your Own Device. *UbiComp 2004 Workshop on Ubiquitous Display Environments*, September, Nottingham, UK.

Bennett, R.E.(2003). *Online Assessment and the comparability of score meaning* (ETS-RM-03-05). Princeton, NJ: Educational Testing Service.

Brown, R. (2012, July). *Microsoft unveils tablet-friendly mice and keyboards (with hands-on).* Retrieved from http://reviews.cnet.com/8301-3134_7-57481956/microsoft-unveils-tablet-friendly-mice-and-keyboards-with-hands-on/

Buxton, W. (1990). The natural language of interaction: A perspective on non-verbal dialogues. In B. Laurel (Ed.), *The art of human-computer interface design* (pp. 405-416). Reading, MA: Addison-Wesley.

Davis, L.L. (2013). *Digital Devices February 2013 Usability Study Report*. Unpublished technical report, Pearson.

Davis, L.L., Orr, A., Kong, X., & Lin, C. (2015). Assessing student writing on tablets. *Educational Assessment, 20,* 180-198.

Davis, L, L. Kong, X. & McBride, Y. (2015).*Device comparability of tablets and computers for assessment purposes*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Davis, L.L, & Strain-Seymour, E. (2013a). *Positioning and ergonomic considerations for tablet assessments.* Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/Positioning.pdf

Davis, L.L, & Strain-Seymour, E. (2013b). *Keyboard interactions for tablet assessments.* Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/Keyboard.pdf

Davis, L.L., Strain-Seymour, E., & Gay, H. (2013). *Testing on tablets: Part II of a series of usability studies on the use of tablets for K-12 assessment programs.* Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/Testing-on-Tablets-Part-II_formatted.pdf

Findlater, L. & Wobbrock, J. O. (2012). Plastic to Pixels: In search of touch-typing touchscreen keyboards. *Interactions*, 19(3), 44-49.

Forlines, C., Wigdor, D., Shen, C., & Balakrishnan, R. (2007, May). *Direct-touch vs. mouse input for tabletop displays.* Paper presented at CHI, San Jose, CA.

Frakes, D. (2013, August). *Buying guide: Find the best iPad keyboard.* Retrieved from http://www.macworld.com/article/1164210/macworld_buying_guide_ipad_keyboards.html

Glader, P. (2013). Amplify's Joel Klein talks tablets, big data, and disappearing textbooks. *THE journal.* Retrieved from: http://thejournal.com/articles/2013/08/08/amplifys-joel-klein-talks-tablets-big-data-and-disappearing-textbooks.aspx

Hall, A.D., Cunningham, J.B., Roache, R.P., & Cox, J.W. (1988). Factors affecting performance using touch-entry systems: Tactual recognition fields and system accuracy. *Journal of Applied Psychology*, 4, 711-720.

Hamdon, N., McKnight, P., McKnight, K., & Arfstrom, K.M. (2013). A review of flipped learning. Retrieved from http://www.flippedlearning.org/cms/lib07/VA01923112/Centricity/Domain/41/LitReview_FlippedLearning.pdf

Heitin, L. (2014, September). Will common core testing platforms impede math tasks? *Education Week*, 34(5). Retrieved from http://www.edweek.org/ew/articles/2014/09/24/05math.h34.html?r=1318807196

Hildaldo, P., & Sato, E. (2014). New technologies to assess English learners. Presented at the annual meeting of the California Educational Research Association. Retrieved from http://images.pearsonassessments.com/images/assets/tell/CERA-Paper-New-Technologies-to-Assess-English-Learners.pdf

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

Hinckley, K. & Wigdor, D. (2011). Input Technologies and Techniques. In Andrew Sears and Julie A. Jacko (eds), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (pp. 161-176). CRC Press.

Holz, C. & Baudisch, P. (2010). *The Generalized Perceived Input Point Model and How to Double Touch Accuracy by Extracting Fingerprints*. Proc. CHI '10, 581-590.

Johnson, D. (2012). Power up! On board with BYOD. *Educational Leadership, 70 (2)*, pages 84-85. Retrieved from: http://www.ascd.org/publications/educational-leadership/oct12/vol70/num02/On-Board-with-BYOD.aspx

Keng, L., Davis, L.L, McBride, Y. & Glaze, R. (2015). *PARCC spring 2014 digital devices comparability research study*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education*, 22(1), 22-37.

Kolen, M. J. (1999). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, 6(2), 73-96.

Lopez, A, & Wolf, M.K. (2013, December). *A Study on the Use of Tablet Computers to Assess English Learners' Language Proficiency*. Paper presented at the annual meeting of the California Educational Research Association, Annaheim, CA.

McCrea, B. (2011). Evolving 1:1: THE Journal. Retrieved from: http://thejournal.com/articles/2011/05/11/evolving-1-to-1.aspx

McCullough, J (2015). *Delivering the national assessment on tablet: Psychometric challenges and opportunities.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*, 449-458.

Meyer, S., Cohen, O., & Nilsen, E. (1994, April). Device comparisons for goal-directed drawing tasks. In *Conference companion on Human factors in computing systems* (pp. 251-252). ACM.

Olsen, J.B. (2014,April). *Score comparability for web and iPad delivered adaptive tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.

Partnership for the assessment of Readiness for College and Careers (2013, February). *Technology Guidelines for PARCC assessments version 2.1 – February 2013 Update.* Retrieved from http://www.parcconline.org/sites/parcc/files/PARCCTechnologyGuidelines2dot1_Feb2013Update.pdf

Pisacreta, D. (2013, June). *Comparison of a test delivered using an iPad versus a laptop computer: Usability study results.* Paper presented at the Council of Chief State School Officers (CCSSO) National Conference on Student Assessment (NCSA), National Harbor, MD.

Pogue, D. (2012, April). *On touch-screens rest your finger by using a stylus.* Retrieved from http://www.nytimes.com/2012/08/02/technology/personaltech/on-touch-screens-rest-your-finger-by-using-a-stylus-state-of-the-art.html?pagewanted=all&_r=0

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment*, 2(6). Retrieved from http://www.jtla.org.

Powers, D.E., & Potenza, M.T. (1996). *Comparability of testing using laptop and desktop computers.* (ETS Report No. RR-96-15) Princeton, NJ: Educational Testing Service.

Prensky, M. (2001). Digital Natives, Digital Immigrants. *On the Horizon* 9(5). Retrieved from: http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf

Raugust, K. (2011)Going mobile in the PreK-12 market. *Simba Information.* Retrieved from: http://www.simbainformation.com/sitemap/product.asp?productid=6055405

Rosin, H. (2013). The touch-screen generation. *The Atlantic.* Retrieved from: http://www.theatlantic.com/magazine/archive/2013/04/the-touch-screen-generation/309250/

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., and Oranje, A. (2005). *Online Assessment in Mathematics and Writing: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series* (NCES 2005–457). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Retrieved from: http://nces.ed.gov/nationsreportcard/pdf/studies/2005457.pdf

Sears, A., & Shneiderman, B. (1991). High precision touchscreens: design strategies and comparisons with a mouse. *International Journal of Man-Machine Studies*, *34*(4), 593-613.

Smarter Balanced Assessment Consortium (SBAC 2013, February). *The Smarter Balanced technology strategy framework and system requirements specifications.* Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Technology-Strategy-Framework-Executive-Summary_2-6-13.pdf

Strain-Seymour, E., Craft, J., Davis, L.L, & Elbom, J. (2013). *Testing on tablets: Part I of a series of usability studies on the use of tablets for K-12 assessment programs.* Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/Testing-on-Tablets-PartI.pdf.

Strain-Seymour, E. & Davis, L.L. (2013). *PARCC device comparability, part I: A qualitative analysis of item types on tablets and computers.*

Strain-Seymour, E. & Brock, B. (2014). *Testing with tablets: What else do you need to know?* Presented at the Council of Chief State School Officers National Conference on Student Assessment, New Orleans, LA.

Stern, J. (2010, May). *The mouse ain't dead…yet: Five of the best mice reviewed.* Retrieved from http://www.engadget.com/2010/05/25/the-mouse-aint-dead-yet-five-of-the-best-mice-reviewed/

Texas Education Agency (2008). *A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests*. Retrieved from http://ritter.tea.state.tx.us/student.assessment/resources/techdigest/Technical_Reports/2008_literature_review_of_comparability_report.pdf.

Tilbrook, D.M. (1976). *A newspaper pagination system*. Department of Computer Science, University of Toronto, Toronto.

van Mantgem, M. (2008). *Tablet PCs in K-12 education.* Retrieved from http://www.iste.org/images/excerpts/TABLET-excerpt.pdf

Vogel, D., & Baudisch, P. (2007, April). Shift: a technique for operating pen-based interfaces using touch. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 657-666). ACM.

Wang, S. (2004). *Online or paper: Does delivery affect results? Administration mode comparability study for Stanford Diagnostic Reading and Mathematics tests.* San Antonio, Texas: Harcourt.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olsen, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, 67(2), 219-238.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olsen, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments. *Educational and Psychological Measurement*, 68(1), 5-24.

Way, W.D. & McClarty, K.L. (2012). Standard setting for computer-based assessments: A summary of mode comparability research and considerations. In *Setting Performance Standards*, ed. G. J. Cizek, 451-466. New York, NY.

Way, W. D., Davis, L. L., & Strain-Seymour, E. (2008). The Validity Case for Assessing Direct Writing by Computer. *A Pearson Assessments & Information White Paper*. Available from *http://images.pearsonassessments.com/images/tmrs/tmrs_rg/TheValidityCaseforOnlineWritingAssessments.pdf?WT.mc_id=TMRS_The_Validity_Case_for_Assessing_Direct*

Way, W.D., Davis, L.L., Keng, L., & Strain-Seymour, E. (in press). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In *Technology in testing: Measurement issues*, ed. F. Drasgow. Vol 2 of the NCME book series.

Winter, P. (2010). *Evaluating the comparability of scores from achievement test variations*. Council of Chief State School Officers: Washington, DC. Retrieved from: http://www.ccsso.org/Documents/2010/Evaluating_the_Comparability_of_Scores_2010.pdf

Yu, L., Lorié, W. & Sewall, L. (2014, April). *Testing on tablets*. Paper presented at the Annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Young, J. G., Trudeau, M., Odell, D., Marinelli, K., Dennerlein, J. T. (2012). *Touch-screen tablet user configurations and case-supported tilt affect head and neck flexion angles. Work* (41), 81-91. Retrieved from: http://iospress.metapress.com/content/x668002xv6211041/fulltext.pdf

NAPLAN Online
Research and Development

Device Effect Study - Field Trial

Contents

1. Executive summary

The National Assessment Program—Literacy and Numeracy (NAPLAN) device effects study was conducted in August 2015 as part of a comprehensive research and development program initiated by the Australian Curriculum, Assessment and Reporting Authority (ACARA) in preparation for the 2017 implementation of the NAPLAN online adaptive testing program. A total of 3602 students across four year levels from 73 different schools participated in two one- hour study sessions to evaluate the impact of different devices (PC, tablet, and tablet with external keyboard) on student testing experience and test performance. In the first session, students responded to test questions in Reading and Numeracy. In the second session, about 20% of students participated in a structured interview as a follow-up to the first session. Both quantitative and qualitative data were collected relative to student testing interactions across the three device conditions. Quantitative data reflected scored student responses to the test items. Using these data, a set of psychometric analyses were conducted to compare student performance across device conditions. Qualitative data were captured through three mechanisms:

- Systematic observations during the test sessions
- Structured interviews after the test sessions
- Follow-up interviews with test administrators

During the systematic observations, test administrators noted both the frequency and severity of issues in the following areas.

- Navigation
- Item specific factors
- Device specific factors
- Domain specific factors
- Ergonomics

These observations were made when observing the class as a whole and again when observing a single student during a "time sampling" period.

The study was also designed to consider whether the performance of students who regularly used a particular device type in the classroom would be less impacted by device effects. This resulted in a 3 (device condition) × 3 (device used in school) set of experimental conditions (9 total conditions) into which schools were recruited for participation. Devices for the study were supplied by the participating schools. The table below shows the range of devices used in the study.

Devices used in the study

Tablet (without external keyboard) (TB)	iPad, iPad Air, iPad mini*, Google Nexus, Samsung tablet
Tablet with external keyboard (TBKB)	Any tablet for which there was an external keyboard, including Microsoft Surface.
Desktop or laptop computer (PC)	Desktop, netbook/laptop (Macbook, Chromebook)

** Note that the iPad mini does not meet ACARA's technical requirements for tablet screen size, but was the device available at some participating schools.*

The test forms used in this study were provided by ACARA and presented to students in their classrooms or a designated testing area, through an online test delivery system procured for the study. Care was taken to include as wide a range of item types as feasible, in order to collect information about the functioning of item design and graphical layout, item and item material interaction, and response types across the assessment domains.

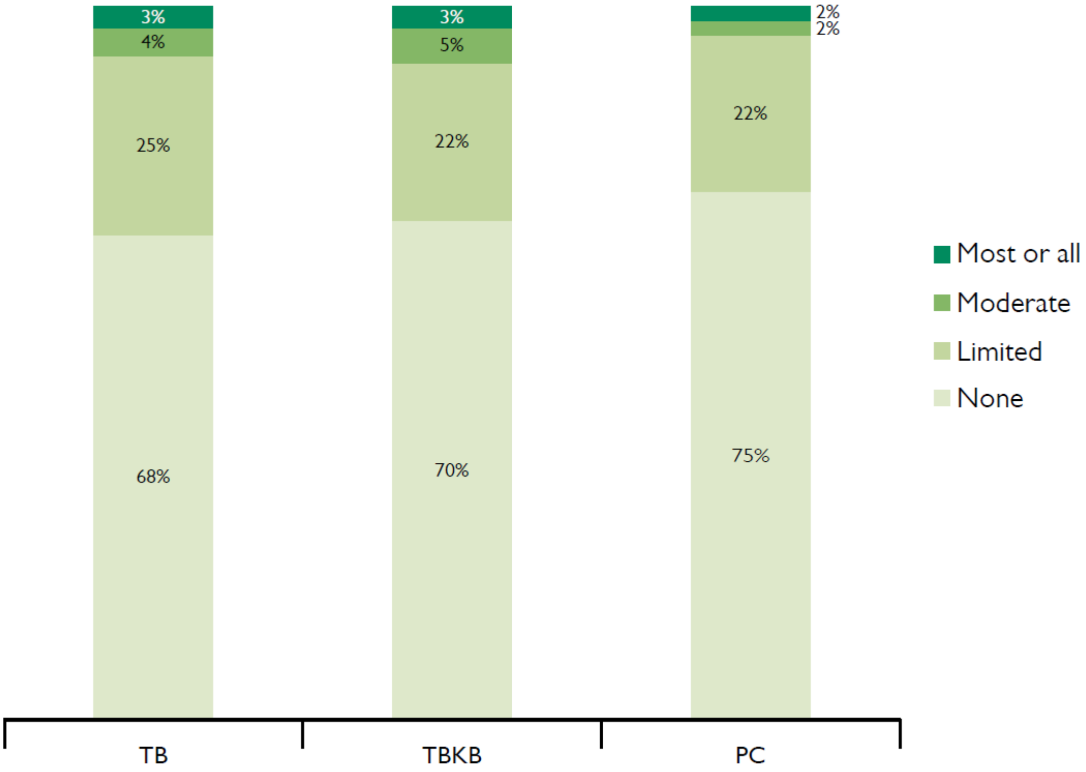
The table below provides a summary of results based on some of the psychometric analyses of Reading and Numeracy. Overall, these results show that there are effectively no differences in student performance across devices for Years 3 and 5. Additionally, these results show some evidence that using the PC was easier than using the TB for students in Years 7 and 9. Importantly, though, statistical analyses show this effect was moderated by student familiarity with the devices, that is, students who were not familiar with using PCs showed no increase in ease of interaction with online tests when doing tests using PC. However, in almost all cases, the TBKB produced comparable results to the PC for students in Years 7 and 9. Additional psychometric analyses provide support for the benefit provided by the external keyboard to student performance, but again confirms that familiarity with the external keyboard, and thus with test device, is important to realise this benefit.

Summary of device effects in psychometric analyses

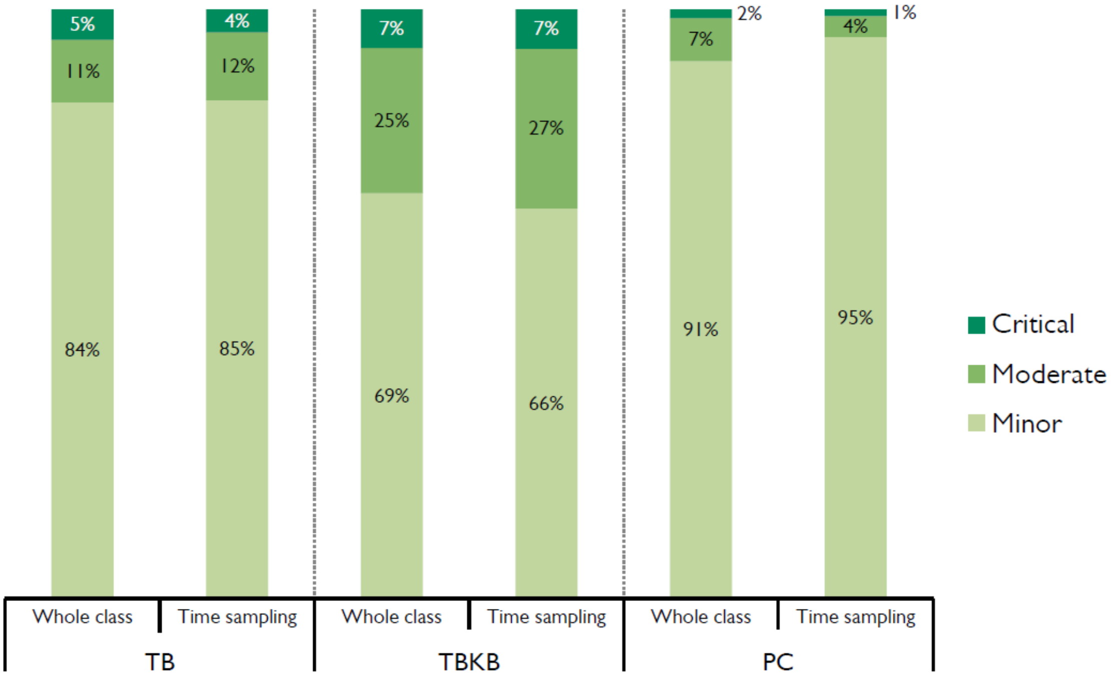
	Year 3	Year 5	Year 7	Year 9
Reading	**No difference**	No difference	PC easier than TB; TBKB same as PC	PC easier than TB; TBKB same as PC
Numeracy	**No difference**	No difference	PC easier than TB; TBKB same as PC	PC easier than TB; TBKB closer to PC

The figures below show the frequency and severity ratings for observed behaviours aggregated across all factors evaluated in the study. Across all device conditions, very few issues were noted by test administrators. This suggests that for the most part students completed their tests without issue. Severity ratings were only provided for the relatively small proportion of observations where issues were observed (approximately 25% to 33% of observations). A relatively small proportion of issues had severity ratings of 'moderate' or 'critical'.

Frequency ratings aggregated across all factors *(Note: Percentages may not add to 100 due to rounding)*



Severity ratings aggregated across all factors *(Note: Percentages may not add to 100 due to rounding)*



Results from both psychometric and qualitative analyses support the conclusion that device effects, when present, are not pervasive but are centered on specific item types, interactions, or interfaces. The salient factors relative to the observation of device effects appear to be:

- use of and familiarity with an external keyboard (especially for older students),
- amount of scrolling required by the test domain interfaces (especially for Reading),
- specific characteristics of drag and drop items; and
- general familiarity with the device, test interface, and item interactions.

Overall results were very similar between classes familiar with and unfamiliar with the testing device. There was, however, some reduction in the observation of issue frequency and severity for classes familiar with the device in areas identified as challenging for students. This suggests that increased exposure to the testing device could ameliorate many of the issues observed.

ACARA is confident that the information provided in this study will allow construction of a NAPLAN online assessment that accounts for any incidental overall device effects assuming student familiarity with the online platform and the device used for the assessment.

2. Introduction

2.1 Background and rationale

In preparation for the 2017 implementation of the NAPLAN (National Assessment Program - Literacy and Numeracy) online testing program, ACARA has developed a comprehensive research agenda that will provide findings on a range of issues and evidence to support the transition of NAPLAN from a paper-based test to a computer-based assessment. An important component of the research agenda is to conduct a device effect study, to investigate the degree of measurement invariance when NAPLAN is delivered across devices, and to identify whether there are variances specific to domains. The information from this study will allow ACARA to develop an online assessment that accounts for any incidental overall device effects assuming student familiarity with the online platform and the device used for the assessment.

The complete study is planned to be conducted in two phases: first, focusing on the measurement invariance of items and testlets and the interaction of students with NAPLAN items in different domains in 2015, and second, in a follow-up study to be conducted in 2016 (when the assessment delivery system that supports the tailored test design will be available) to investigate the device mode effect in context of adaptive testing focusing on the invariance of student performance. The following research report relates to the 2015 Phase 1 study only.

Information on students' experience completing a range of NAPLAN tests within an online environment was gathered using systematic observation and structured interview methods to collect qualitative data on the interaction and engagement of students with online tests delivered on different devices. Additionally, student response data were captured and used to conduct psychometric evaluations of differences in performance across devices.

The devices used in the 2015 study were divided into three categories which are in general use in most Australian schools: personal computers (PCs); tablets (TBs); and tablets with external keyboards (TBKBs). Students from Years 3, 5, 7 and 9 were selected to participate in the study, with a total of 3602 participants. Students undertook a combination of tests from the following domains: Numeracy, Reading and Spelling. The tests were delivered using an online test delivery system procured for this study.

2.2 Summary of literature review

In preparation for conducting the Device Effect Study, ACARA commissioned a literature review for the purpose of aiding their understanding of the impact of testing device (especially laptops and tablets) on the performance of items and students in online assessments. The literature review was intended to guide and support the development of protocols to be used for systematic observations and structured interviews. Additionally, as relevant, results of this study will be addressed in relationship to the research summarised in the literature review.

The literature review situated research on device effects as part of the larger domain of score comparability as defined by Bennett as "the commonality of score meaning across testing conditions including delivery modes, computer platforms, and scoring presentation" (Bennett, 2003). The author discussed the form factors of different devices and suggested how these differences might relate to expected comparability (Way, Davis, Keng, & Strain-Seymour, 2015). Devices which have similar form factors (meaning devices have similar physical properties and students interact with them through similar input/output mechanisms) can be expected to have a relatively high degree of comparability, while devices with dissimilar form factors (meaning devices have disparate physical properties and students interact with them using different mechanisms for input/output) can be expected to have a lower degree of comparability.

Within this framework, the author describes the specific features of touch-screen devices (like tablets) that might influence the comparability of scores between tablet- and computer-delivered tests. These include screen orientation, pinch-and-zoom, device position, and the use of the finger as a pointer. Research suggests that students generally prefer to use the tablet in landscape orientation and most typically work with the tablet positioned flat on a table surface while leaning over it (Davis, 2013; Davis, Strain- Seymour, & Gay, 2013). Additionally, students were observed using pinch-and-zoom (though somewhat sparingly) when they were having difficulty viewing the test content. The size of objects relative to the students' finger is an important consideration when designing content for tablets. When students are using their finger to select and move objects, the degree of precision is less than it is with a mouse and consequently objects need to be larger. Despite these differences, research in this area has primarily found non-statistically significant and/or non-practically significant effects at the total score level (Keng, Davis, McBride, & Glaze, 2015; Davis, Kong, & McBride, 2015; Olsen, 2014).

Special attention is given to discussion of onscreen and external keyboards as input mechanisms for text. External keyboards allow for three states of hand positioning (fingers are off the keys; fingers are resting on the keys; fingers are depressing a key), whereas onscreen keyboards only allow for two states of hand positioning (fingers are off the keys; fingers are depressing a key; Findlater & Wobbrock, 2012). This lack of a resting state with an onscreen keyboard creates challenges for the use of conventional keyboarding techniques. The compact size of an onscreen keyboard also means that students are working in a more constrained space when reaching fingers to select keys. Several studies suggest that students tend to write less under these circumstances than they would with an external keyboard (Davis, Strain-Seymour, and Gay, 2013; Strain-Seymour, Craft, Davis, & Elbom, 2013). However, Davis, Orr, Kong, and Lin (2015) did not find differences in either the length or quality of student writing when using an onscreen versus an external keyboard in a study conducted with Grade 5 and high school students (Grades 10 and 11).

The literature review concluded that assessment content can be delivered comparably across devices with appropriate attention given to user interface and technology requirements. Specifically, the review concluded that device effects, when present, tend to be small and isolated to specific item types or interactions. Furthermore, students themselves appear to be fairly robust to device effects and adapt quickly to new methods of interaction. The literature review, however, also cautioned that the use of touch-screen devices for assessment introduces a series of new and important considerations for test developers. Specifically, the Reading domain should be given careful consideration as it typically requires students to read significant additional material, such as a stimulus text.

Moreover, individual item interactions need to be carefully reviewed on each type of device to be used for testing. This should allow any potential usability issues or areas of confusion for students to be identified and addressed in subsequent research and development activities.

The literature review also made specific recommendations for ACARA in terms of:

- the use of onscreen versus external keyboards for tablets,
- the interface for drag-and-drop items,
- the amount of scrolling required for reading passages,
- device positioning and ergonomics,
- the use of touch-screen specific features; and
- the audio functionality within the Spelling test.

With regard to keyboards, specific recommendations were made to consider the amount of text entry required by different item types; the age level of students using the keyboards, the type of external keyboard used, and the amount of experience students had using the keyboards provided.

A specific focus on drag and drop items was suggested given the variety of different types of drag and drop interactions as well as previously observed usability challenges with some items of this type.

For reading passages, recommendations were made to observe both the amount of scrolling needed (especially for longer passages) and the mechanisms students attempted to use for scrolling (finger, mouse, etc.).

For device positioning and ergonomics, recommendations were made to observe the general position and use of devices to evaluate the impact this appeared to have on student fatigue or eye strain.

With regard to touch-screen specific features, recommendations were made to observe the degree to which students rotated their tablets, used pinch-and-zoom, etc. and the circumstances under which these behaviours tended to occur.

Finally, with regard to audio functionality it was noted that research has not yet been conducted into the effects of this factor across devices. In the absence of research in this area recommendations were made to note any challenges students had with controlling the volume, using the onscreen controls, and using features like audio-playback.

ACARA used the literature review findings and recommendations to finalise the design of the 2015 device effect study presented in this report. To address the lack of research findings regarding the impact of familiarity with device on performance in online tests, ACARA designed the device effect study as a counter design where familiarity of students with the test device was explicitly manipulated and controlled in the study. In order to explicitly address issues regarding text entry when tests were taken using tablets, tablets with external keyboards were included as a separate device effect condition and the familiarity of students with test devices was also explicitly investigated. In addition, ACARA's test managers used the literature review to inform and guide selection of items and construction of tests used in the study.

3. Methods

3.1 Overview of methods

Data for this study were collected from a total of 3602 students in 73 schools across four Year levels (3, 5, 7, and 9). The sample was a convenience sample, constructed in collaboration between ACARA and the Test Administration Authorities (TAAs). TAAs were asked to provide a selection of schools across school systems and geographical locations. Thus the final sample contained a good cross-section of schools at the national level.

Students participated in two one-hour sessions. In the first session, they responded to test questions in Reading, Numeracy, and Spelling). In the second session, about 20% of students participated in a structured interview as a follow-up to the first session. Both qualitative and quantitative data were collected relative to student testing interactions across three device conditions.

- Tablet (without external keyboard) (TB)
- Tablet with external keyboard (TBKB)
- Desktop or laptop computer (PC)

The study was additionally designed to include a second independent variable related to device familiarity in order to evaluate the hypothesis that if students regularly used a device in the classroom, their performance would be less impacted by device effects. This resulted in a 3 (device condition) × 3 (device used in school) set of experimental conditions (9 total conditions) into which schools were recruited for participation.

Qualitative data was captured through three mechanisms:

- Systematic observations during the test sessions
- Structured interviews after the test sessions
- Follow-up interviews with test administrators

Quantitative data reflected scored student responses to the test items. Using these data, a set of psychometric analyses were conducted to compare student performance across device conditions.

3.2 Participants

Recruitment for the study was targeted to achieve participation of 36 classes per year level divided across 9 experimental conditions for a total of 4 classes per experimental condition per year level. Assuming 25 students per class, this would provide 100 students per experimental condition at each year level. In order to make most efficient use of school participation, two classes were recruited from each school when possible. During the recruitment process, however, it became apparent that devices were not all equally used in classrooms or available for testing. Specifically it was challenging to find schools using tablet devices consistently across classes and most challenging to find schools using tablets with external keyboards. This made it difficult to achieve the targeted distribution of classes and students fully crossed by experimental conditions.

Priority was therefore given to identifying classes to participate in the hardest to fill experimental conditions even if both classes from a given school could not participate in those conditions. Frequently, a school might have enough tablet devices for one class, but not two classes. The end result of this was an oversampling of classes in the computer (PC) conditions to achieve a minimum level of participation in the tablet (TB) and tablet with external keyboard (TBKB) conditions.

Table 1 provides a summary of student (and class) participation by experimental condition across year levels. A total of 73 schools across Australia were recruited to participate in the study. The schools were selected from different jurisdictions and sectors. The sample included both primary and secondary schools. While the majority of the sample was drawn from schools in metropolitan and regional areas, some remote schools participated in the study. Jurisdictions were asked to nominate schools with a good spread of student backgrounds and also to ensure that the selected schools participated in the study. Overall, a total of 148 classes, consisting of 38 Year 3 classes, 45 Year 5 classes, 37 Year 7 classes, and 28 Year 9 classes, were involved. Note that for 34 classes more than one device was represented within a single class. These “hybrid” classes are, therefore, represented twice in the table (once for each device used in study testing). Despite efforts to recruit classes for all experimental conditions, the number of classes who had prior experience using tablets with external keyboards was very small and the TBKB/PC condition was dropped from Year levels 5, 7, and 9.

Table 1
Summary of student (and class) participation

Device used in Classroom	Device used for Testing	Year 3	Year 5	Year 7	Year 9	Total
PC	PC	285 (14)	275 (14)	451 (21)	345 (14)	1356 (63)
	TB	92 (7)	168 (9)	84 (6)	29 (3)	373 (25)
	TBKB	118 (5)	155 (6)	31 (2)	20 (2)	324 (15)
TB	PC	148 (6)	104 (4)	92 (4)	86 (4)	430 (18)
	TB	100 (4)	143 (6)	82 (3)	95 (2)	420 (15)
	TBKB	78 (3)	130 (5)	78 (3)	51 (2)	337 (13)
TBKB	PC	24 (1)	–	–	–	24 (1)
	TB	40 (2)	52 (2)	28 (2)	13 (1)	133 (7)
	TBKB	27 (1)	83 (3)	30 (1)	65 (3)	205 (8)
Total		912 (43)	1110 (49)	876 (42)	704 (31)	3602 (165)

A subsample of classes (about 20% of all participating classes) was selected to participate in a structured interview during the second hour of the study in order to provide feedback about experiences during testing. Table 2 provides a summary of class participation in the structured interview process.

Table 2
Structured interview participation

Device used in Classroom	Device used for Testing	Year 3	Year 5	Year 7	Year 9	Total
PC	PC	1	3	5	4	13
	TB	0	2	2	0	4
	TBKB	1	1	0	0	2
TB	PC	1	1	0	1	3
	TB	0	1	2	0	3
	TBKB	1	0	0	1	2
TBKB	PC	1	–	–	–	1
	TB	0	0	0	1	1
	TBKB	0	1	0	1	2
Total		5	9	9	8	31

3.3 Materials

Devices

Devices for the study were supplied by the participating schools. In most cases, devices were assigned to individual students. In a few cases, devices (typically desktop computers) were located within technology labs at the school. Schools were assigned to device conditions based upon the available technology. As indicated in Table 3, a range of different devices were used within each study condition.

Table 3
Devices used in the study

Tablet (without external keyboard) (TB)	iPad, iPad Air, iPad mini*, Google Nexus, Samsung tablet
Tablet with external keyboard (TBKB)	Any tablet for which there was an external keyboard, including Microsoft Surface.
Desktop or laptop computer (PC)	Desktop, netbook/laptop (Macbook, Chromebook)

** Note that the iPad mini does not meet ACARA's technical requirements for tablet screen size, but was the device available at some participating schools.*

The PCs used in the study were mainly laptop devices with case-based external keyboards. In some cases the laptops were connected to a full size keyboard and external mouse. Few schools had purchased external keyboards for their tablet devices, so external keyboards were provided for 19 of the 35 classes participating in the tablet with external keyboard (TBKB) condition. These keyboards were full sized, ergonomically designed, and contained keys which provided good feedback to users.

Moreover, the keyboards were wireless, thereby allowing the tablet to be rotated if necessary. Some of the tablets used in the study were equipped with their own keyboards set within the device case. The use of external keyboards with tablets seemed to create an additional layer of technical challenge. In some situations the external keyboards did not work or stopped working during the test administration. Additionally, the external keyboard did not always over-ride the onscreen keyboard thus complicating the process of entering a response. A small number of students experienced accidental disconnection issues and needed to have their keyboards reconnected. Some of these issues were caused by students accidentally pressing the wireless function key. It is likely that many of these issues could be overcome if external keyboards were part of the school's standard equipment as they would be optimised to work with the school's devices and students would have sufficient experience to work with them appropriately.

Test forms

The tests used in this study were provided by ACARA and presented to students in their classrooms or a designated testing area, through an online test delivery system procured for the study. Table 4 shows the total number of items and distribution of item types reflected on the Reading and Numeracy test forms for each year level. Note that care was taken to include as wide a range of items as feasible, in order to collect information about the functioning of item design and graphical layout, item and item material interaction, and response types across the two assessment domains. Definitions and examples of each item type are presented in Appendices A and B.

Table 4
Test form summary for Reading and Numeracy

	Reading				Numeracy			
Item type	Year 3	Year 5	Year 7	Year 9	Year 3	Year 5	Year 7	Year 9
Multiple choice	15	20	18	25	9	6	9	5
Multiple select	3	0	4	3	1	1	6	5
Hot spot	0	2	0	0	0	2	0	1
Drag and drop	6	2	5	1	7	11	11	12
Inline choice	0	0	1	0	4	4	1	1
Short response text entry	0	0	2	0	2	3	5	8

Extended response text entry	2	2	2	0	0	0	0	0
Composite	0	0	0	3	1	1	0	0
Total number of items	26	26	32	32	24	28	32	32
Number of passage texts	5	5	9	6	N/A	N/A	N/A	N/A

The Spelling test for all year levels consisted of nine short answer response type items requiring the entry of a single word. Each word was read to students in the context of a sentence via an audio feature within the online test delivery system. Students wore headphones during this test and could adjust the volume within the program. They were also able to replay the item as many times as they wanted.

Prior to taking the tests in each domain, students completed a practice test. The practice test consisted of nine questions which reflected most of the item types that students would encounter in the Reading, Numeracy and Spelling tests. This test was designed to be completed without any assistance from the test administrator.

3.4 Test administrators

Test administrators were responsible for the critical activity of recording device and item related events as they transpired in the classroom. With this in mind, the test administration contractor initiated a recruitment drive with the aim of employing test administrators with experience working in a classroom setting with primary and secondary level students. Candidates were selected on the basis of their previous experience in:

- exam invigilation,
- working with primary and secondary level students,
- data collection, particularly interviewing,
- trialling new systems and associated research activities; and
- report writing.

In recognition of the study's goal to test students across a range of devices, successful candidates were also required to:

- have strong IT skills,
- be familiar with a range of devices; and
- have IT troubleshooting experience.

Qualified test administrators attended a training session conducted by the test administration contractor. During this session the design, structure and aims of the research activity, and its data collection methods, were explained.

Training focused on:

- understanding the approved systematic observation and structured interview protocols,
- the accurate recording of information collected during observations and interviews,
- the use of the test administration contractor's data management software to record these data; and
- the use of a recording device provided for the interview sessions.

3.5 Process

Students participated in two one-hour sessions. In the first session, they responded to test questions in two of three domain areas (Reading, Numeracy, and Spelling). In the second session, about 20% of students participated in a structured interview as a follow-up to the first session. Table 5 summarises the combinations of tests and interviews delivered in schools.

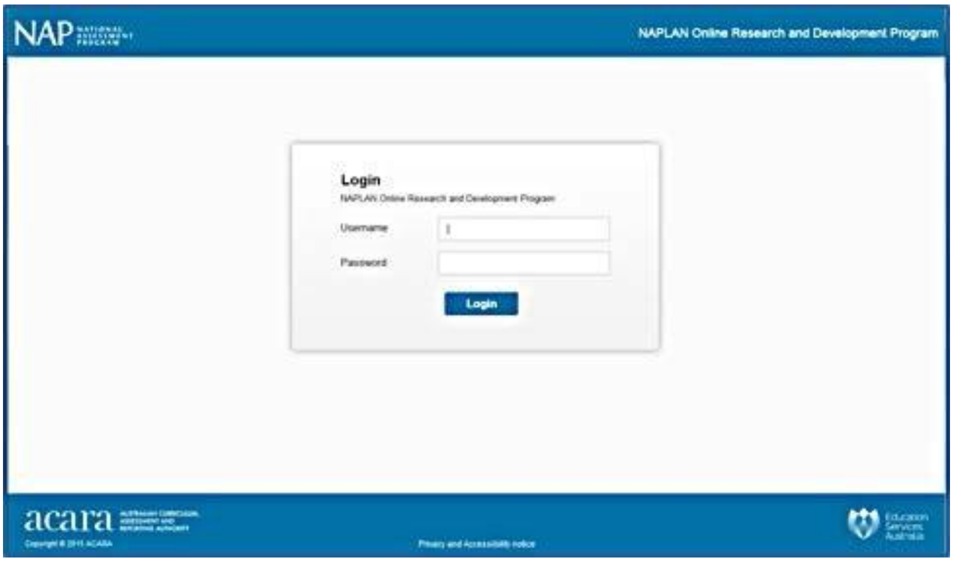
Table 5
Summary of test sessions

	Session 1		Session 2	
Display condition	Domain areas	Session time	Domain areas	Session time
Test Form 1	Numeracy and Reading	1 hour	Structured interview	1 hour
Test Form 2	Reading and Spelling	1 hour	Structured interview	1 hour
Test Form 3	Spelling and Numeracy	1 hour	Structured interview	1 hour

Test administrators were required to arrive at the study site ahead of the session starting time to ensure that the room was set up for the tests. However, devices were only available in advance of the scheduled testing session for sites using stand-alone desktop computers. In most cases, devices were not available for set-up until students arrived for the study with either their own devices or with a portable device provided by the school.

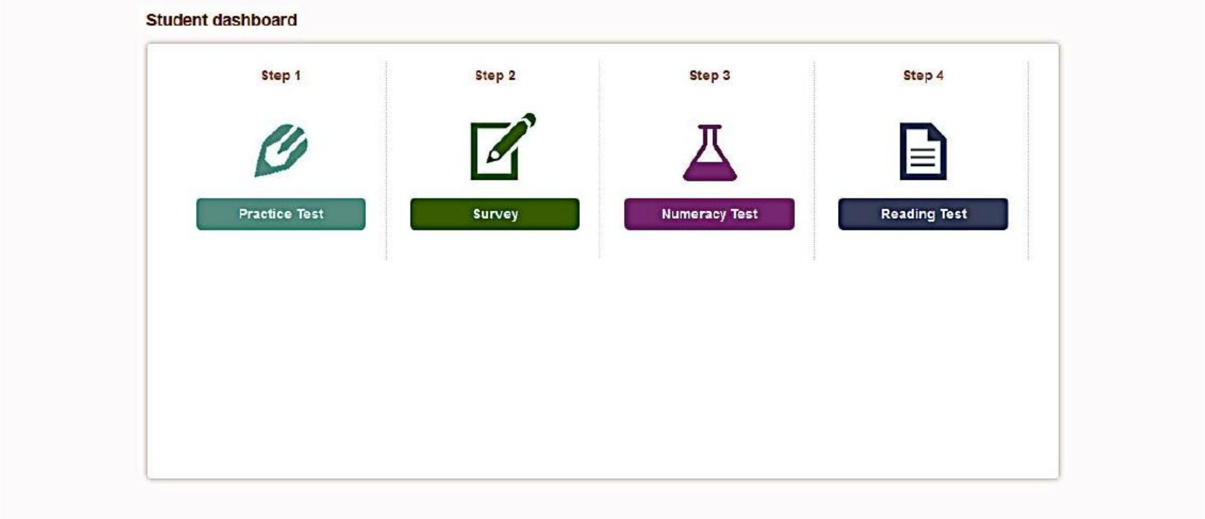
Upon arrival, students were briefed on the purpose of the study, told how much time had been allocated to each test, and that they would be told when this time was almost finished. Students were then directed to access the testing software by entering a URL on their devices. After entering the URL to access the test delivery system, students were presented with the login screen (see Figure 1) into which they needed to enter the unique ID number and password which had been provided to them on a separate information sheet.

Figure 1
Login screen


NAP NATIONAL ASSESSMENT PROGRAM
NAPLAN Online Research and Development Program
Login
NAPLAN Online Research and Development Program
Username
Password
Login
acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY
Copyright © 2015 ACARA
Privacy and Accessibility notice
Education Services Australia

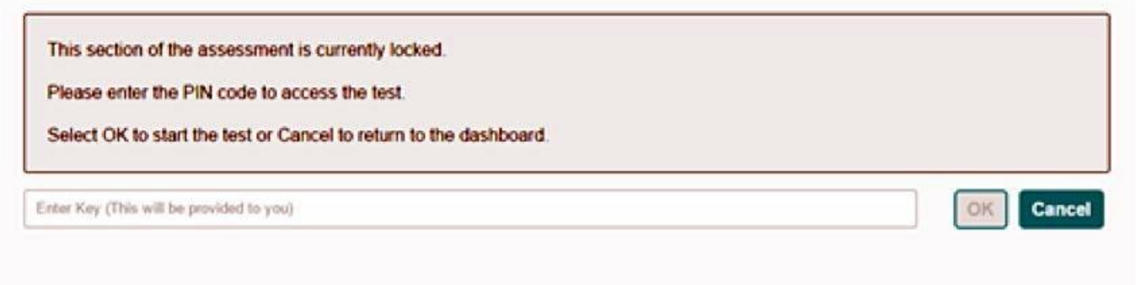
After logging in, students were presented with a dashboard containing a series of icons (see Figure 2). These provided entry points to the following activities; a survey on device usage, the practice test, and the suite of tests allocated to that student ID. Students were directed to complete the survey and the practice test before beginning the domain area tests.

Figure 2
Student dashboard



When students had completed the practice test, they were presented with a screen (see Figure 3) which prompted them to enter a four digit PIN for the first domain area test. Once they had completed and submitted the first test, students were prompted to enter the PIN for the second test. The PIN to be used for each test was written on the board by the test administrator so that students could move onto the next test as soon as they had completed the first.

Figure 3
Test login screen


This section of the assessment is currently locked.

Please enter the PIN code to access the test.

Select OK to start the test or Cancel to return to the dashboard.

3.6 Data collection

During the test sessions, qualitative data were captured by the test administration contractor through systematic observations and structured interviews. Research protocols for both the observations and interviews were developed by the test administration contractor and were reviewed and approved by ACARA. Additionally, the test administration contractor conducted follow-up interviews with test administrators.

Systematic observation

The research protocols for the systematic observation included two different methods for collecting the data:

‘event-frequency’ and ‘time-sampling’. The event-frequency method required test administrators to observe as many students as possible for two thirds of the allotted time. The time-sampling method required test administrators to identify one or two students and observe them, and only them, for one third of the allocated observation time. Two separate protocols were developed for use at each year level - one for the event-frequency observation and one for the time-sampling observation. In total, test administrators completed four protocols for each testing session - two for each of the two domains tested. An example of the systematic observation protocol for the event sampling observation is presented in Appendix C.

The protocols outlined the following areas for observation:

- Navigation
- Item specific factors
- Device specific factors
- Domain specific factors
- Ergonomics.

The event frequency protocol contained fields for test administrators to use during the whole class observation period to make notes about specific observed behaviours. These fields included:

- Question number - as test administrators observed student interactions they recorded the relevant question number.
- Frequency - test administrators used hash marks to record each individual occurrence of the observed behaviour.
- Tally - test administrators totalled the number of hash marks for each observed behaviour.
- Researcher’s notes - test administrators made notes about the observed behaviours for later reference.

At the end of each protocol, test administrators were provided with a space to record any additional comments. In addition, test administrators were asked to summarise their observations using a frequency rating scale and a severity rating scale. These ratings were to be based on the tallies and notes made during the observation period. The frequency rating scale was defined as follows:

- ‘1’: No issues observed.
- ‘2’: Issues observed for only a limited number of students (e.g. around one-quarter).
- ‘3’: Issues observed for a moderate number of students (e.g. around one-half).
- ‘4’: Issues observed for most or all students.

The severity rating scale was defined as follows:

- ‘1’: Minor – causes some hesitation or slight irritation.
- ‘2’: Moderate – causes occasional task failure for some users; causes delays and moderate irritation.
- ‘3’: Critical – leads to task failure; causes user extreme irritation.

Note that severity ratings should only have been assigned if the frequency rating for a behavior was a ‘2’, ‘3’, or ‘4’. If a frequency rating of ‘1’ (no issues observed) was assigned, then there would be no severity rating for that behavior.

The protocol for the time sampling observation was very similar to that for the event frequency observation. The same four fields (question number, frequency, tally, and researcher's notes) were provided for each behavior. The main difference was that test administrators were instructed to record an observed behavior as many times as it occurred during the observation period for the selected student rather than for the class as a whole. As with the event frequency protocol, test administrators had a space provided at the end of the time sampling protocol to record any additional comments. For the time sampling observation, test administrators were asked to summarise their observations using the severity rating scale. No summary frequency rating was provided for the time sampling observation as the frequency rating scale was defined relative to the proportion of students (limited, moderate, most or all) who were observed having the issue.

Hard copy forms were used by test administrators to record their observations. This information was later entered electronically into the test administration contractor's data management software designed for the study.

Structured interview

To facilitate discussion within groups and between the interviewer and students, structured interviews consisted of a series of open-ended questions, which, apart from some domain-specific questions, were the same for each testing domain. During the interviews test administrators displayed hard copy images of the test items to facilitate student recall.

The protocols outlined the following areas for discussion:

- Navigation
- Item specific factors
- Device specific factors
- Domain specific factors
- Ergonomics.

The interviewer would prompt students by asking questions such as:

- How easy was it for you to move around the test?
- Tell me how you felt about scrolling in the passage.
- Did you know what to do with this type of question?
- Was there anything confusing?

The interviewer recorded student responses in the open spaces provided on the protocol next to each topic area. Additionally, permission had been requested to record interviews, but this was not granted by all schools.

Follow-up interviews with test administrators

The test administration contractor project management staff conducted exit interviews with all test administrators at the conclusion of the testing period. The aim of these interviews was to gather additional information about the data collection process and test administrators' perspectives of students' interactions with device types. This information was entered into the test administration contractor's data management software to supplement the observation and interview data.

3.7 Data analyses

Qualitative data analyses

Information from the systematic observations, structured interviews, and exit interviews with test administrators was reviewed and summarised for this report using standard qualitative analysis techniques to identify relevant themes across classes and device conditions. Additionally, test administrator ratings of frequency and severity from each observed behavior have been summarised in charts to support these themes.

Psychometric data analyses

Student response data from the device effect study were collected and scored to support a set of psychometric comparisons across device conditions. Two primary sets of analyses were conducted for the Reading, Numeracy, and Spelling domains - one focusing on performance of items which involved the calibration of Rasch Model item difficulty values through the software program Conquest (Wu, Adams, & Wilson, 2007), and a second focusing on performance of students using a Generalised Linear Mixed Model (GLMM) approach. Details of the methodology used with each of these analyses are provided below.

Rasch Model item analysis

Student response data from all device conditions were calibrated simultaneously using the Conquest software program under three separate models. In the first model, item difficulty was freely estimated without regard to the testing device. This served as the baseline model and assumes no device effect. In the second model, the testing device was included as a test-level facet. A parameter was estimated which described the deviation from the baseline model item difficulty for each device condition. This model assumes a constant effect of device across all items within a test. The third model included testing the device as a facet at the item-level and allowed a unique deviation from the baseline model item difficulty for each item within each device condition. The three models are summarised as follows, where “TestC” provides the facet for device condition.

- Model 1: item
- Model 2: item + TestC
- Model 3: item + item * TestC

Note that for the Rasch analyses an attempt was made to include the full set of 3 × 3 experimental conditions but the models would not converge. Therefore only testing device is represented in this analysis. Unlike the GLMM analysis where a single data matrix was constructed for analysis, all three models were run separately across Year levels 3, 5, 7, and 9 for each domain (Reading, Numeracy, and Spelling). This yielded a total of 12 different sets of Rasch Model outputs. For Numeracy Year level 9, Model 3 (item level device effects) would not converge so results from that model are excluded for that domain/year level. Note that while results for spelling converged under each of the three models, these results were ultimately not interpretable because auto-correct and auto-complete were enabled across many student devices which invalidated the performance data results.

GLMM analysis

For each year level, student responses were submitted to a Generalised Linear Mixed Model (GLMM) analysis combining the domains of Numeracy, Reading, and Spelling. This was possible because of the pairing of test administrations across the three domains (Numeracy, Reading and Spelling) that allowed for the construction of a data matrix connecting responses of all students across the domains. In contrast to Rasch Model analyses which have the key purpose to establish a unidimensional measurement scale, the purpose of GLMM analyses are to estimate the impact of independent variables on a dependent variable and thus the constructed data matrix was suitable for use in GLMM analyses. Additional advantage of GLMM is that it provides a method to model the impact of independent variables as fixed factors, such as item and device types and as random factors such as selection of students and students used in this study.

GLMM also provides for a straightforward comparison of models which have a different composition of random factors, fixed factors, and the interaction between these factors. To that end in each year level a set of models of increased complexities were analysed and compared. Following such model fit comparison, a single model was used to analyse the effect of fixed factors in more detail.

Where more than one model showed a statistically significant fit to the data, the model with the most parsimonious composition of fixed and random factors was selected to complete the GLMM analyses.

- Model 1: assumes that experimental condition (in this case the full 3x3 set of experimental conditions) is the only independent variable and that students and items are random factors.
- Model 2: introduces item interaction type (drag, click, etc.) as an independent variable in the model, in addition to experimental condition.
- Model 3: allows for an interaction between the independent variables of experimental condition and item interaction type.
- Model 4: adds domain as a third independent variable.
- Model 5: takes the nesting of items within domain into account as a random factor.
- Model 6: allows for all possible two-way interactions between the three independent variables (experimental condition, item interaction type, and domain).

4. Results

In the following sections, results from the psychometric analyses are presented first followed by results from the qualitative analyses. The psychometric analyses describe in which areas and to what degree device effects impacted student performance in this study. The qualitative analyses support a detailed explanation of how underlying issues may have contributed to differences in student performance across devices.

Additionally, the qualitative analyses describe challenges students had in interacting with the test content due to the software interface, the device, or the combination of these two factors. These challenges may not always have been significant enough to impact student performance but, nonetheless, should be understood and addressed when possible.

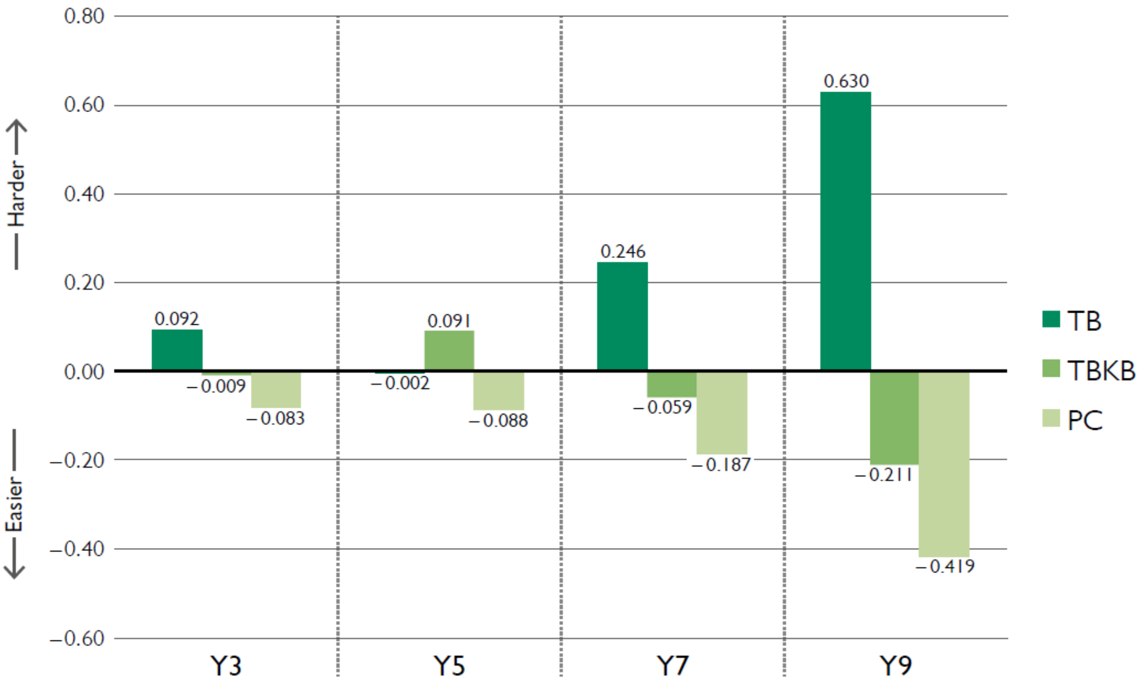
4.1 Results of psychometric analyses

Rasch analyses

Figures 4 and 5 show the test level device effects (Model 2 of the Rasch analysis) for Reading and Numeracy. There are no statistically or practically significant differences in student test performance across testing devices for Year levels 3 and 5. Larger differences were observed in student test performance for Year levels 7 and 9 in both domain areas. For Reading, the tablet condition (TB) was more difficult than the PC condition at both Years 7 and 9 (p<.05; effect size of 0.44 logits for Year 7 and 0.69 logits for Year 9). However, the results for the TBKB condition are not statistically or practically different from those from the PC condition at either year level.

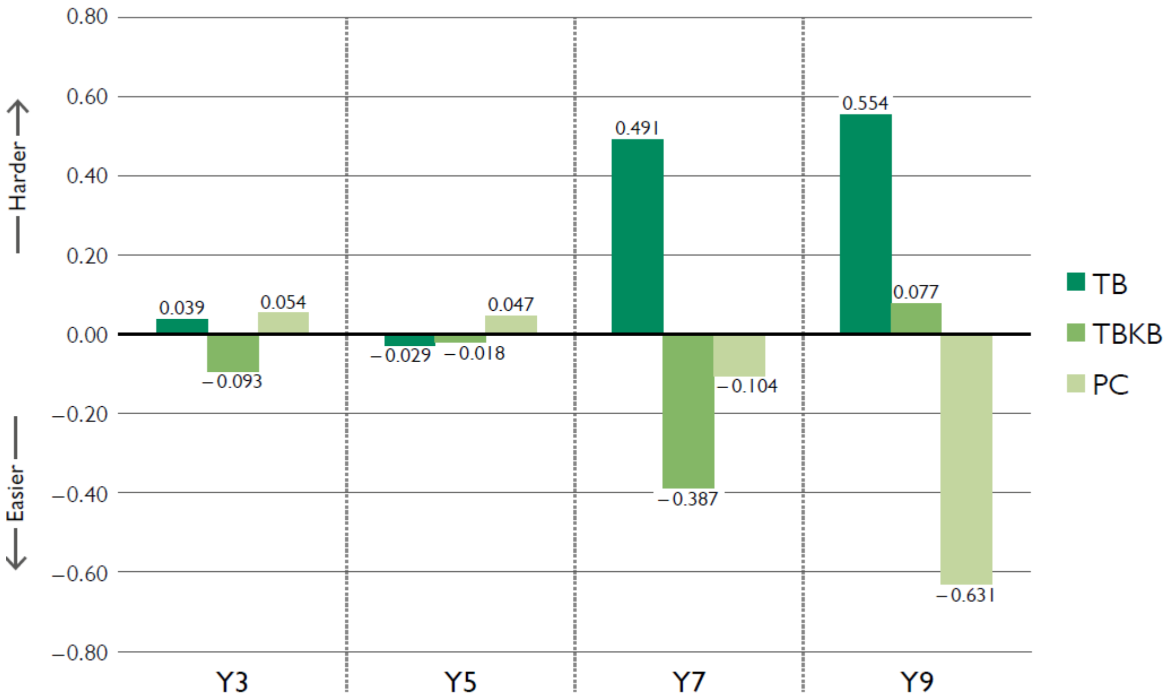
Figure 4

Device effect by year level (Reading)



For Numeracy, the tablet condition was again more difficult than the PC for both year levels (p<.05; effect size of 0.23 logits for Year 7 and 0.73 logits for Year 9). The TBKB condition was not statistically or practically different from the PC condition for Year 7. While the TBKB condition was closer in difficulty to the PC condition than the TB condition at Year 9, the difference between TBKB and PC was still statistically significant for Year 9 (p<.05 with an effect size of 0.48).These results suggest that the external keyboard may benefit older students on the Reading and Numeracy tests.

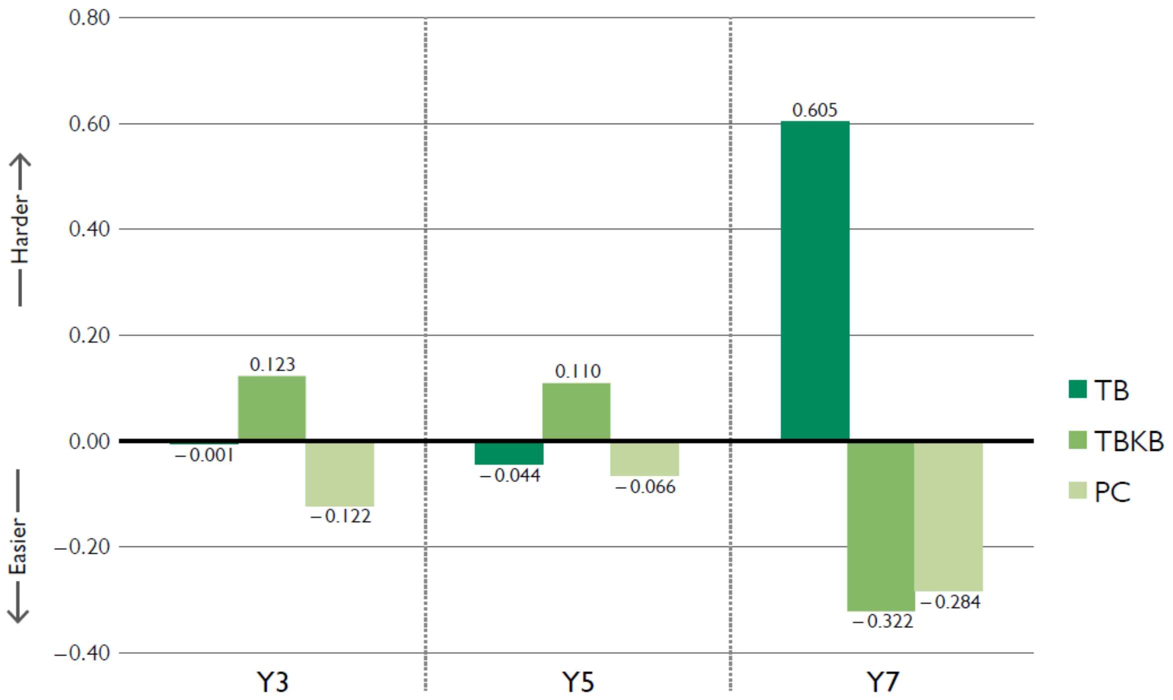
Figure 5
Device effect by year level (Numeracy)



Evaluation of Rasch device effects for different item types revealed that using different devices had little to no impact on student performance for item types requiring click interactions (multiple choice, multiple select, and hot spot). Some very small effects were observed for drop down items (inline choice) in the Year 3 and 5 Numeracy tests (see Figure 6) favouring students testing on PC over those testing with TB or TBKB.

Figure 6

Device effect for inline choice items (Numeracy) *(Note: Year 7 results reflect 1 item and should not be over-interpreted)*



Larger differences in performance across devices were observed for both drag and drop and text entry items. As seen in Figures 7 and 8, the Rasch device effect for the drag and drop items appeared to favour students testing on TB in earlier Year levels (3 and 5) and students testing on PC or TBKB in later Year levels (7 and 9) - the one exception being Reading Year 3 where effects appeared to favour students testing on PC rather than TB. Results from the Rasch item analysis suggest that drag and drop items were particularly difficult for Year 7 Numeracy TB users and Year 5 Reading TBKB users.

Figure 7

Device effect for drag and drop items (Reading) *(Note: Year 9 results reflect 1 item and should not be over-interpreted)*

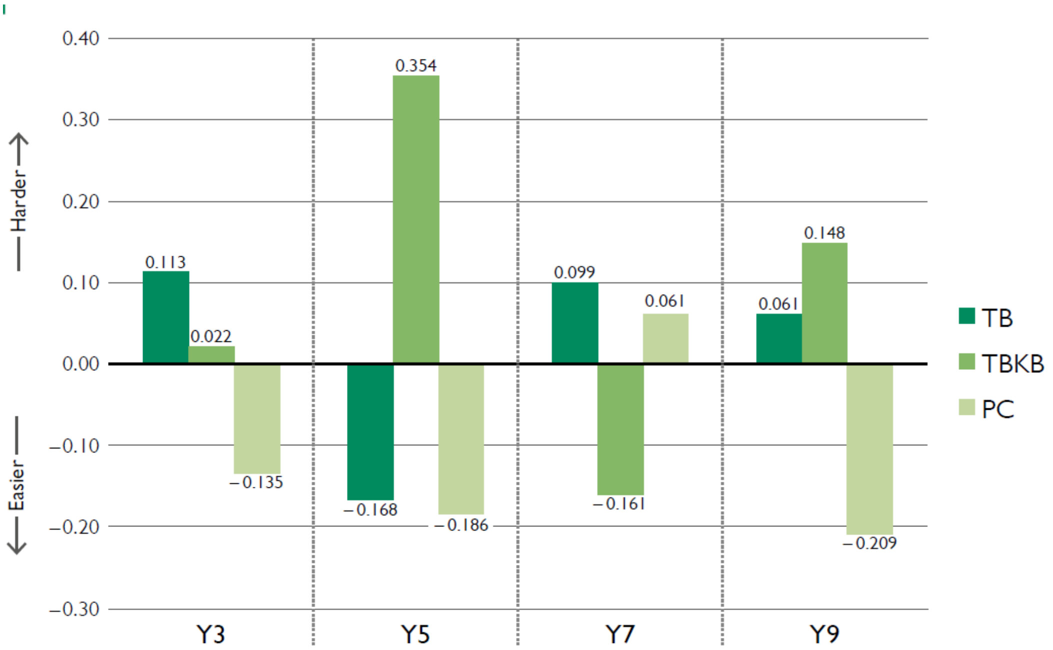
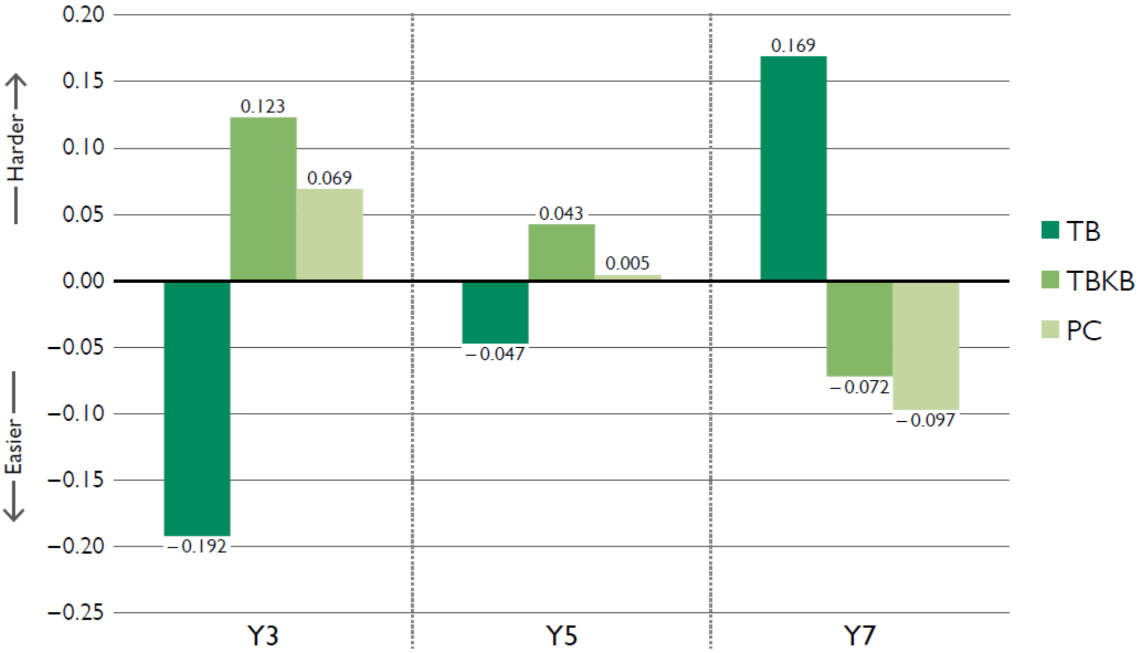


Figure 8

Device effect for drag and drop items (Numeracy)



As seen in Figures 9 and 10, the Rasch device effect for text entry items was the most consistent across year levels and domains. These items appeared most difficult for TB users in all year levels and in both the Reading and Numeracy tests. They were easiest for TBKB users in all but the Year 7 Reading test (where results were very similar across devices).

Figure 9

Device effect for text entry items (Reading)

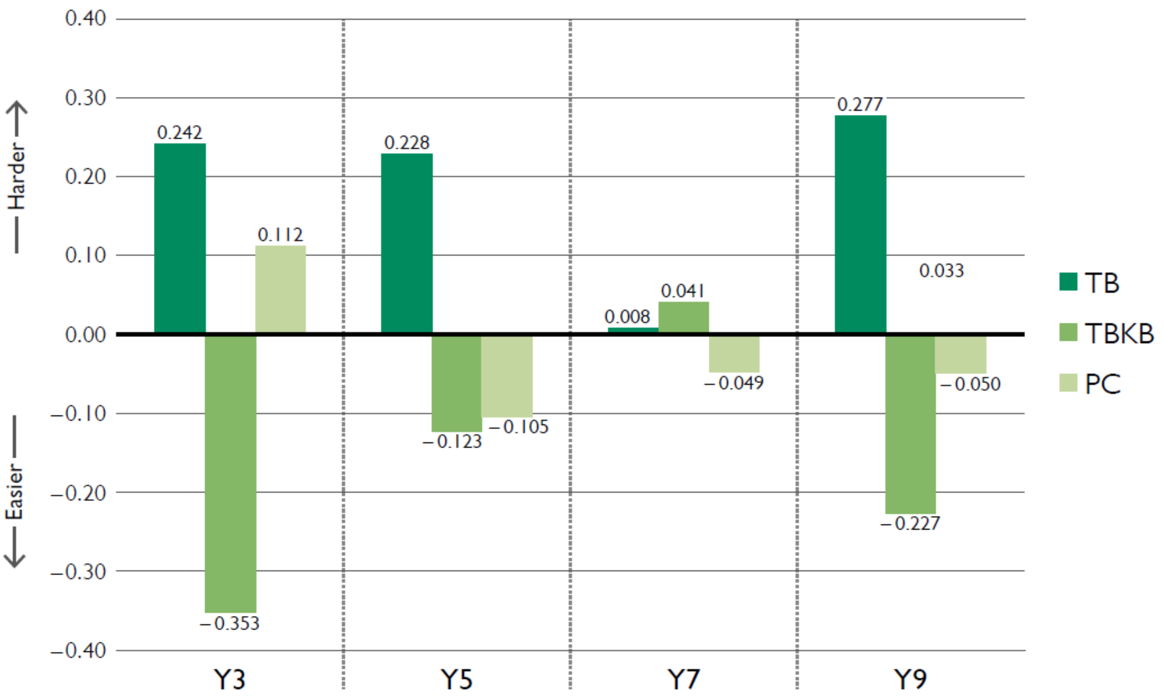
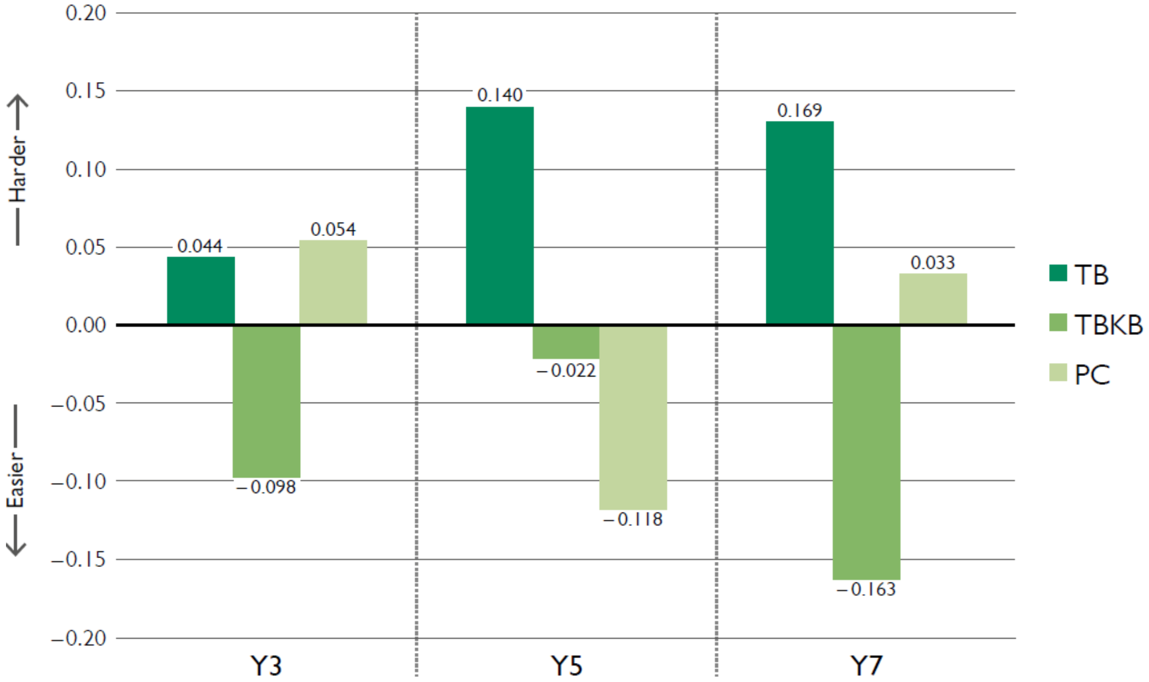


Figure 10
Device effect for text entry items (Numeracy)



GLMM analyses

The results from the GLMM analysis showed that Model 2 (which includes both experimental condition and item type as independent variables) was the most parsimonious model that fit the data for Year levels 3, 5, and 7 (chi-square of 8.91, p=.03 for Year 3; chi-square of 9.92 for Year 5, p=.02; chi-square of 27.5, p<.001 for Year 7). Model 3 (which allows an interaction between experimental condition and item type) was the most parsimonious model that fit the data for Year level 9 (chi-square of 82.2, p <.001).

To investigate and compare different conditions within and across each of the fixed factors, treatment coding was used in presented GLMM analyses. The multiple choice and TB-TB conditions were set as the base against which percentage correct in all other conditions were compared. For Year 3, the analyses showed that there were no statistically significant differences in performance across experimental conditions. The only significant difference observed at Year 3 was with regard to item type - text entry items were more difficult than multiple choice items across all devices (p<.01).

For Year 5, the analyses showed a significant main effect for experimental condition. Students who had previous experience with tablets with external keyboards and who also used them in the study (TBKB- TBKB condition) had a greater likelihood of answering items correctly than students in the TB-TB condition (p=.002). However, this same benefit was not seen for students who did not have previous experience with the external keyboards (students in the PC-TBKB and TB-TBKB conditions). Additionally, students who had previous experience with tablets with external keyboards and did not use them in the study (TBKB-TB condition), had a lower likelihood of answering items correctly than students in the TB-TB condition (p=.05). Although these effects are negligible to small in terms of the effect size (95% confidence interval of-0.76 to -0.17 logits and 0.003 to 0.74 logits, respectively), these results indicate that students with previous experience with tablet and keyboard benefit from having the keyboard in the test situation.

Results from the Year 7 analysis show a significant main effect for both item type and experimental condition. With regard to item type, text entry and drag and drop items were difficult than multiple choice items across all devices (p<.001 and p-.02, respectively). With regard to experimental condition, students who did not have experience the external keyboard with tablets, but were asked to use it in the study (TB-TBKB condition) had significantly lower correct response rates than students in the TB-TB condition (p<.001; small to moderate effect with 95% confidence interval ranging from 0.39 to 1.09 logits). Students in the PC-PC condition also had significantly lower correct response rates than students in the TB-TB condition (p=.04), but these results were of negligible effect size (95% confidence interval ranging from 0.006 to 0.50 logits).

For Year 9, there was a significant interaction between item type and experimental condition. Drop down items (inline choice) were easier for students in some experimental conditions (PC-PC, p<.001; PC- TBKB, p=.04; TB-PC, p<.001) than they were for students in the TB-TB condition. Drag and drop items were easier for students in the TBKB-TBKB condition than they were for students in the TB-TB condition (p=.005). Text entry items were more difficult for students in the PC-TB condition than in the TB-TB condition (p=.04) and easier for students in the PC-TBKB condition than in the TB-TB condition (p=.04). Both of these results suggest that there may be some benefit to having the external keyboard for drag and drop and text entry items.

In summary, the GLMM analysis provides some evidence that having the external keyboard with the tablet can benefit student performance if students are familiar with that configuration. Students in the TBKB-TBKB condition outperformed students in the TB-TB condition overall at year 5 and for drag and drop and text entry items at year 9. Additionally students in the PC-TBKB condition outperformed students in the TB-TB condition for year 9. For all but this last finding, familiarity with the external keyboard was important to be able to realise an increase in performance as students in the PC-TBKB and TB-TBKB conditions did not see this same benefit. Outside of these findings for the external keyboard, effects of device on student performance (overall, by domain, or by item type at other year levels) were largely absent from the results.

Summary of psychometric results

Table 6 provides a high level summary of the results of the Rasch analyses for Reading and Numeracy relative to device effects across domains and year levels. Overall, these results show that there are effectively no differences in student performance across devices for Years 3 and 5. Additionally, these results show fairly consistent evidence that the PC was easier than the TB for students in Years 7 and 9. However, in almost all cases, the TBKB produced comparable results to the PC for students in Years 7 and 9. Importantly, the GMM analyses which investigated the impact of deliberate manipulation and control of familiarity with the test devices unambiguously showed that the familiarity with the test device, rather than the device itself, is the key factor impacting student interaction with online tests. The GLMM analyses provide support for the benefit provided by the external keyboard to student performance, but again suggest that familiarity with the external keyboard is important to be able to realise this benefit.

Table 6

Summary of device effects in psychometric analyses

	Year 3	Year 5	Year 7	Year 9
Reading	**No difference**	No difference	PC easier than TB; TBKB same as PC	PC easier than TB; TBKB same as PC
Numeracy	**No difference**	No difference	PC easier than TB; TBKB same as PC	PC easier than TB; TBKB closer to PC

4.2 Results of qualitative analyses

Figure 11 shows the frequency ratings for behaviours observed during the event frequency (whole class) observation period, aggregated across all factors evaluated in the study. Across all device conditions, very few issues were noted by test administrators. A frequency rating of 'no issues observed' was given for 68-75% of the observations. A frequency rating of 'moderate' or 'most or all' was only given for 4-8% of the observations. This suggests that for the most part students completed their tests without issue. There were slightly more issues observed for TB and TBKB conditions than for the PC condition, but these differences are relatively small.

Figure 12 shows the severity ratings for behaviours observed during both the event frequency (whole class) and time sampling (individual student) observation periods, aggregated across all factors evaluated in the study. Severity ratings were only provided for the relatively small proportion of observations where issues were observed (approximately 25% to 33% of observations). In some cases, this was a very small number of observations (as few as four) and, as such, severity ratings should not be over-interpreted. A relatively small proportion of issues had severity ratings of 'moderate' or 'critical'. Severity ratings were slightly higher for TB than for PC and highest for TBKB.

Figure 11
Frequency ratings aggregated across all factors *(Note: Percentages may not add to 100 due to rounding)*

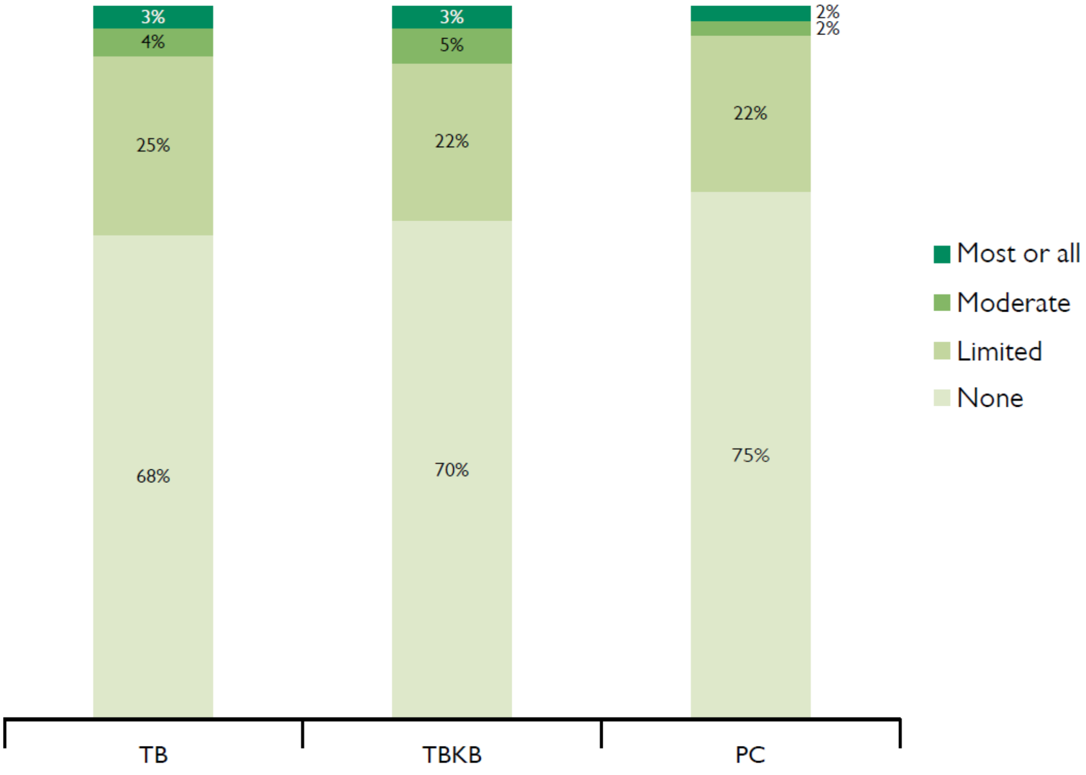
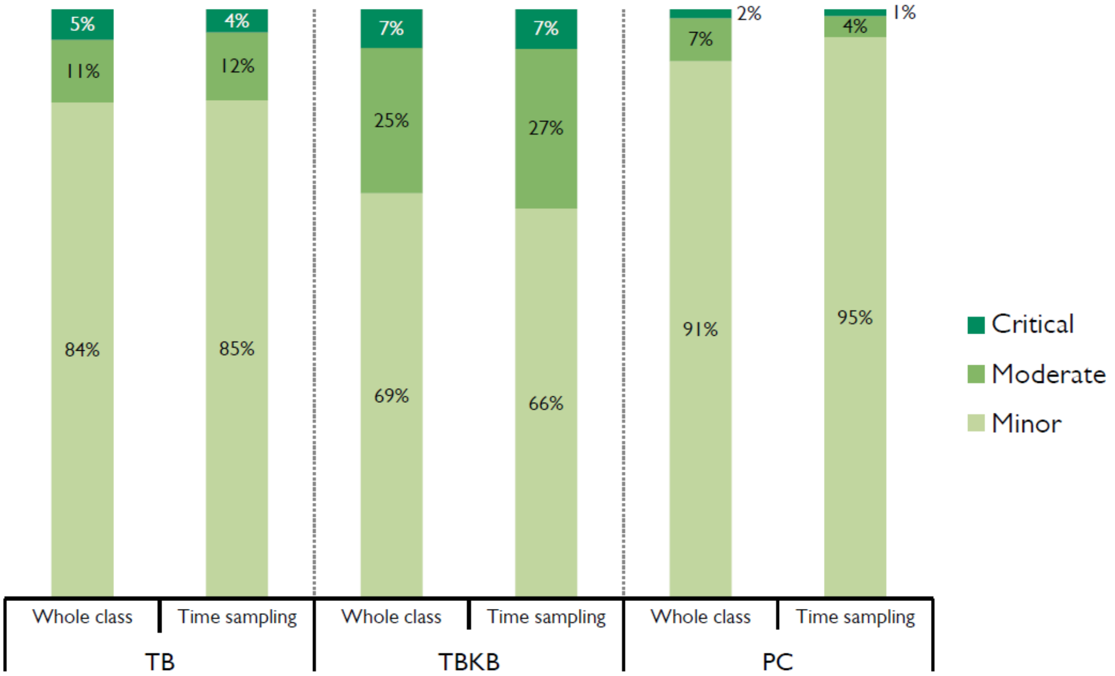


Figure 12
Severity ratings aggregated across all factors *(Note: Percentages may not add to 100 due to rounding)*

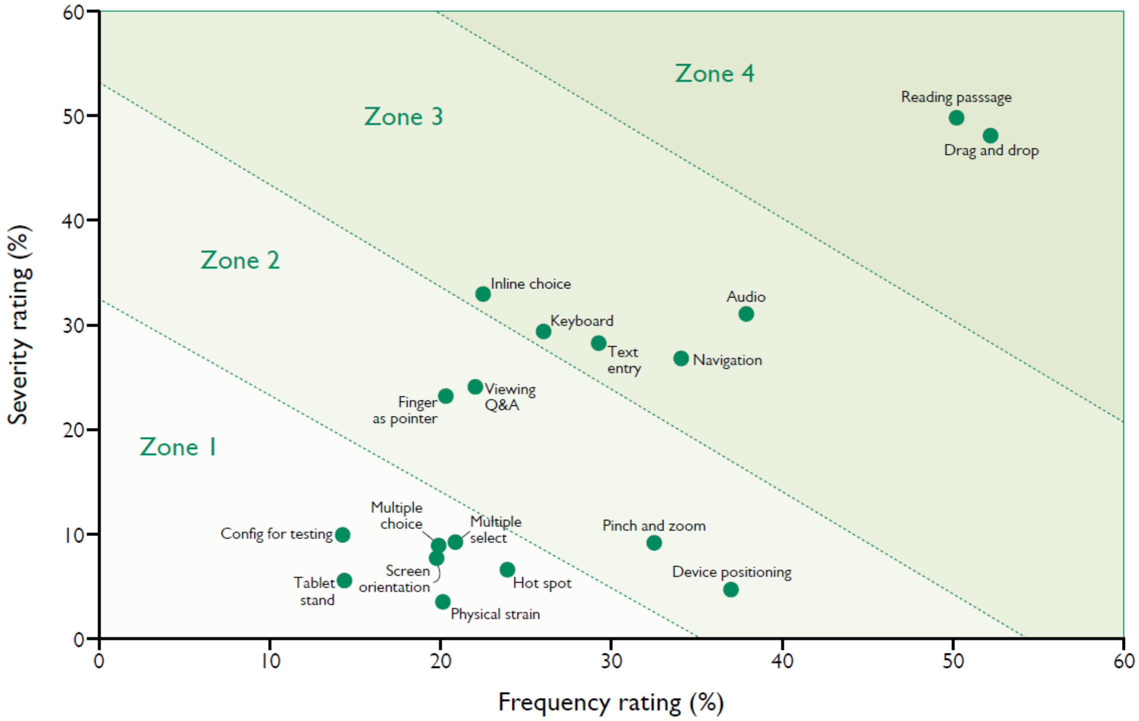


Across all device conditions, frequency ratings indicated issues were observed less than 1/3 of the time. When an issue was observed, a rating of ‘minor’ severity was given more than 2/3 of the time. Taken together, this means that issues which were of moderate or critical severity were observed for approximately 11% of the observations (1/3 frequency * 1/3 severity). While patterns in both the frequency and severity of issues reported do show some differences across devices, the overall small prevalence and magnitude of issues suggests that device effects, when present, were not widespread or large, but instead were isolated to functionality with specific item types or device features.

Figure 13 overlays frequency and severity ratings across device conditions and year levels for all behaviours rated during the event sampling (whole class) observation period. To simplify interpretation, the ratings in this chart are dichotomised relative to each rating scale. The frequency ratings shown in this figure reflect the percent of observations with a rating of ‘2-limited’, ‘3-moderate’, or ‘4-most or all’. In other words, the issue is observed (rating of 2, 3 or 4) rather than not observed (rating of 1). The severity ratings shown in this figure reflect the percent of observations with a rating of ‘2-moderate’ or ‘3-critical’. In other words, the issue severity is not minor (rating of 2 or 3) rather than minor (rating of 1).

Figure 13

Clusters for behavioural observations: Ratings averaged across devices and year levels



In reviewing the data in Figure 13, four zones emerge relative to the observed behaviours. Zone 1 reflects the lowest frequency and severity ratings. Zone 4 reflects the highest frequency and observation ratings.

See Appendix C for an example of the observation protocol which provides a detailed description of each of the behaviours reflected in this figure. Note that many of these behaviours are inter-related such that some behaviours may have been observed at an elevated level of frequency or severity in response to other behaviours. For example, challenges with drag and drop items or reading passages may have led students to demonstrate greater use of pinch and zoom. The remaining sections within the qualitative results will describe the issues observed within Zones 3 and 4 in greater detail and will reflect on differences across devices for these behaviours as well as how different device interactions may have been used by students in response to these issues. Specifically, the following topics will be reviewed in more detail.

- Navigation between items
- Reading passages and interface
- Writing and keyboards
- Spelling audio items and headphones
- Results by item type
  - Click interactions (multiple choice, multiple select, and hot spot)
  - Inline choice (drop down) items
  - Drag and drop items
  - Text entry items

- General device interactions

While these discussions may focus on the challenges students encountered and the differences across devices, it should be noted that the overall results demonstrate that students were largely successful in engaging with the test content and that device effects were minimal.

Navigation between items

Figures 14 and 15 show the frequency and severity ratings across all domains and year levels for navigation within the test. This includes use of the next and previous buttons to move from one item to the next, use of the item number toolbar to move from one item to the next, use of the flag and review buttons, and any navigation that exited students from the testing application. Issues with navigation were more frequently noted in the TBKB condition than either the TB or PC conditions with 23% of test administrators indicating that navigation issues occurred for a moderate number of students or for most or all students. The severity of issues with navigation was also greater with TBKB than with the TB or PC conditions with approximately half of the issues observed with the TBKB condition classified as moderate or critical compared with 20% or fewer of the issues in the other device conditions.

Figure 14
Frequency ratings for navigation *(Note: Percentages may not add to 100 due to rounding)*

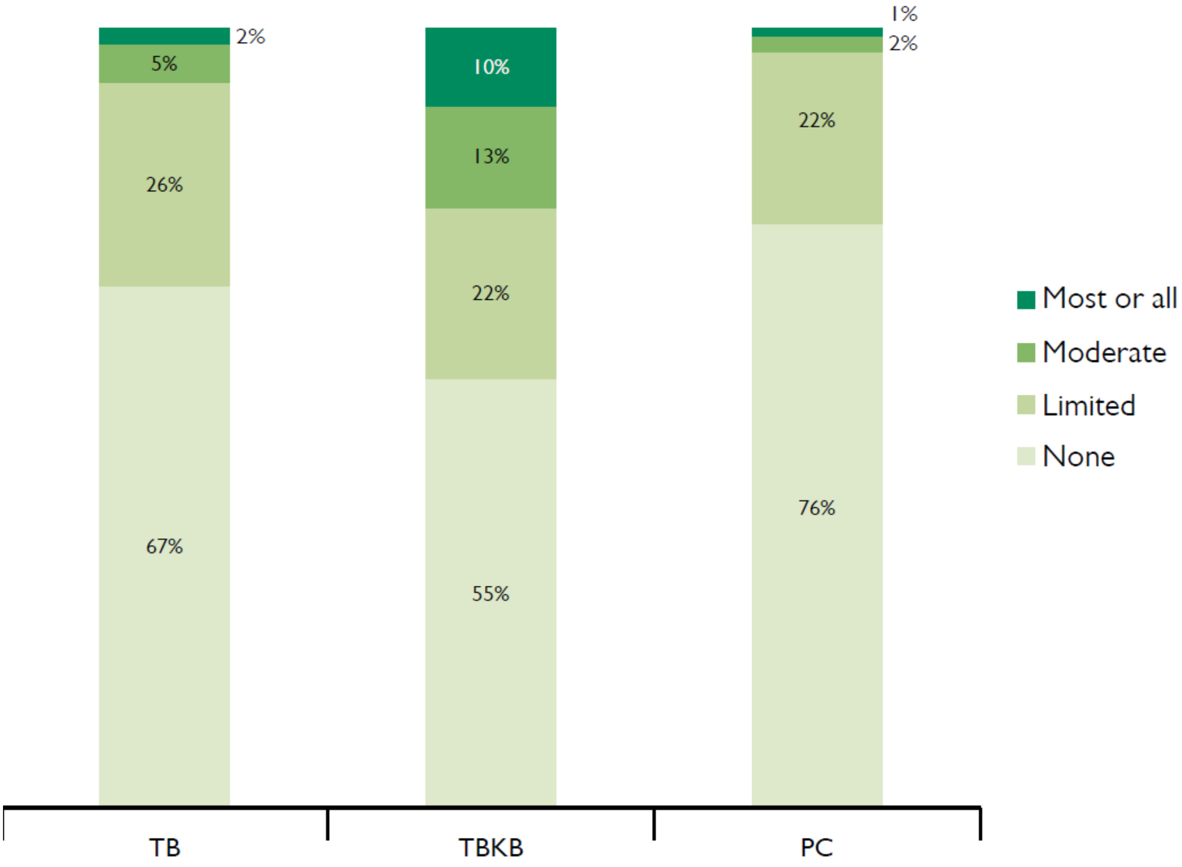
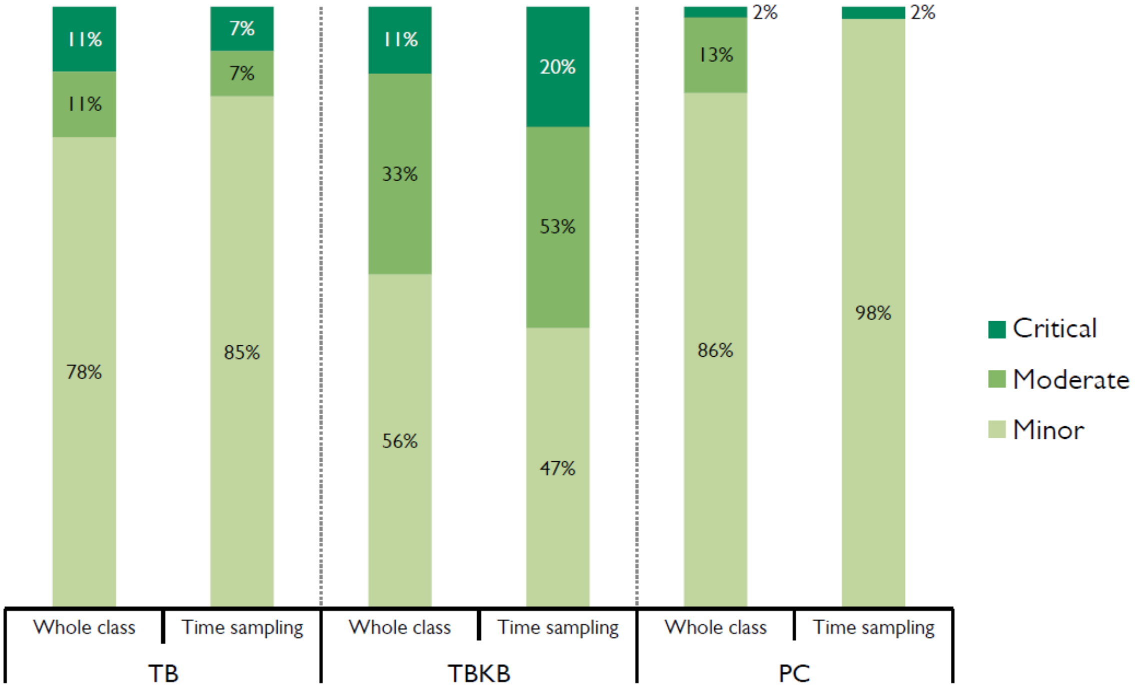
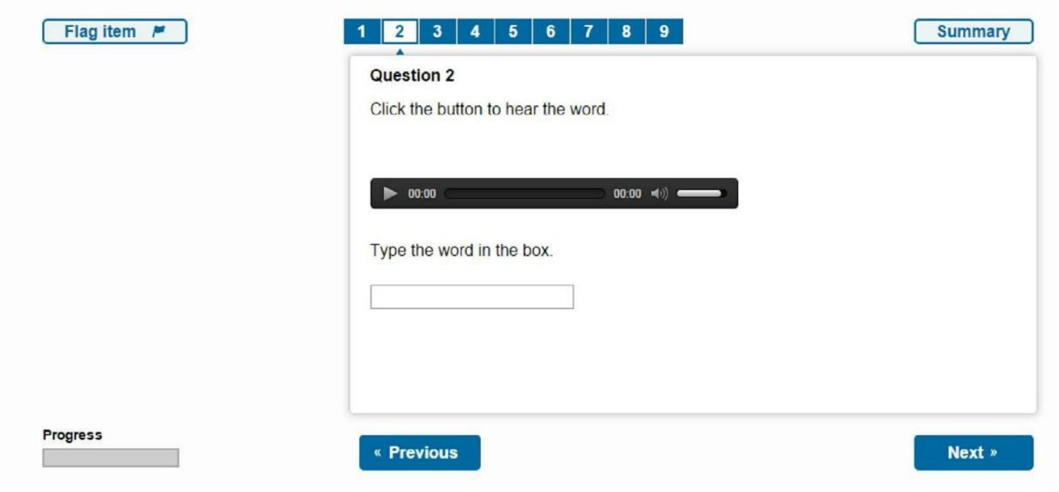


Figure 15
Severity ratings for navigation *(Note: Percentages may not add to 100 due to rounding)*



Across all devices, most students were able to problem-solve many of the issues encountered with navigation during the tests. Students were observed to develop and employ strategies which worked best for them - though there were frequently tradeoffs of time, efficiency, and visibility of content. Although most students said that navigating the tests was not confusing, some students said that "it wasn't as easy to use as it could be". Figure 16 provides a sample screenshot (shown with a Spelling item) of the test interface showing the navigation controls. The Reading, Numeracy, and Spelling domains each contained the same general controls for navigating between items - the 'Next' and 'Previous' buttons at the bottom of the screen as well as the item tool bar at the top of the screen (this allowed navigation to a specific item).

Figure 16
Navigation controls within the test interface



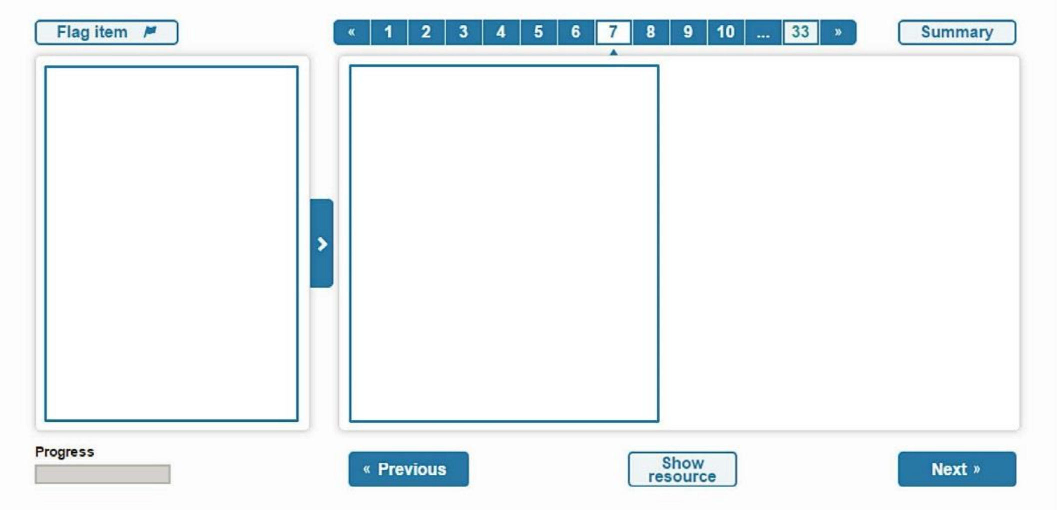
Question 2

Click the button to hear the word.

Type the word in the box.

The Reading interface, as shown in Figure 17, had a panel to the left of the item which contained the text of the reading passage. Students could use the blue expansion tab or the 'Show resource' button to expand the passage text across the screen for easier viewing. When the passage text was expanded the 'Show resource' button was replaced with a 'Hide resource' button.

Figure 17
Reading interface with passage panel minimised



Across all devices students found the 'Next' and 'Previous' buttons easy to use, but thought that the item toolbar functionality was not intuitive. In many cases students only began using this toolbar after its function was explained to them. Test administrators observed that TB and TBKB tended to use both the item toolbar and the 'Next' and 'Previous' buttons to navigate between items, while PC users generally used only the 'Next' and 'Previous' buttons. PC users tended to use the item toolbar only when they wanted to return to specific questions.

Although presenting a more intuitive navigation control than the item toolbar, access to the 'Next' and 'Previous' buttons was a concern for students within the Numeracy and Reading tests. For Numeracy, the large size of some items meant that the 'Next' and 'Previous' buttons were not always visible on the screen without scrolling. Some TB and TBKB users resolved this by using pinch and zoom (zooming out so as to see everything) or by rotating the tablet from landscape to portrait orientation to allow more vertical space. However, the trade off for either of these solutions was a reduction in the size of the item text and in the response features of some item types. Additionally, test administrators observed that some Year 3 and Year 5 students had difficulty controlling the degree of zoom, which frequently exacerbated the navigation issues or created different issues in viewing the items. Test administrators also noted that some Year 3 students in the TB and TBKB conditions would accidentally select the 'Next' and 'Previous' buttons while trying to scroll to see the item content.

Students encountered the same issues relative to the 'Next' and 'Previous' buttons with the Reading tests as they did with the Numeracy tests. However, in most cases it was the reading passage that caused the scrolling rather than the item itself. The reading passage did not scroll independently of the item as a whole. Therefore, when the passage text was lengthy, the 'Next' and 'Previous' buttons were pushed "down the page" such that they were only visible when students had scrolled to the end of the passage.

This frequently resulted in students losing visibility of the item itself - scrolling so much that the item was no longer visible in the right hand panel. Therefore, when students did select 'Next,' they had no visual feedback as to whether or not they had moved on to the next item. The right hand panel appeared blank before and still appeared blank. This seemed to cause confusion, as students would have to scroll back up to confirm if they had successfully moved on to the next item. As with the Numeracy test, students in the TB and TBKB conditions attempted the same solutions of zooming out to see the entire item and rotating the tablet from landscape to portrait. However, given the larger degree of scrolling with the reading passages, these approaches tended to be less successful and the reduction to item and passage text was greater. As with Numeracy, Year 7 and 9 students generally found the zoom function to be helpful in managing the complexities of this navigation, but some Year 3 and 5 students encountered issues with controlling the zoom in these navigation processes.

Navigation issues were largely not observed on the Spelling test because the item content was relatively compact and students did not need to scroll, zoom, or change tablet orientation to view all the components of the item and response area.

Reading passages

Figures 18 and 19 show the frequency and severity ratings across all year levels for observed issues with reading passages. This includes:

- difficulty seeing the item and passage at the same time,
- difficulty using the expansion tab to minimise or expand the reading passage,
- difficulty using the 'Show/Hide resource' button to minimise or expand the reading passage,
- difficulty scrolling to read the passage and/or excessive scrolling; and
- the need to enlarge or zoom in on the passage text.

Across all devices, issues with reading passages were observed at a higher rate (in 43-54% of the observations) than with the overall frequency rating across behaviors (see Figure 18). Additionally, issues with reading passages were noted somewhat more frequently in the TB and TBKB conditions than in the PC conditions. This suggests that the reading passage interface was challenging for students across all device conditions, but may have been more challenging for students in the TB and TBKB conditions. The severity rating for observed issues was higher for the TBKB condition than for the TB or PC conditions.

Figure 18
Frequency ratings for reading passages *(Note: Percentages may not add to 100 due to rounding)*

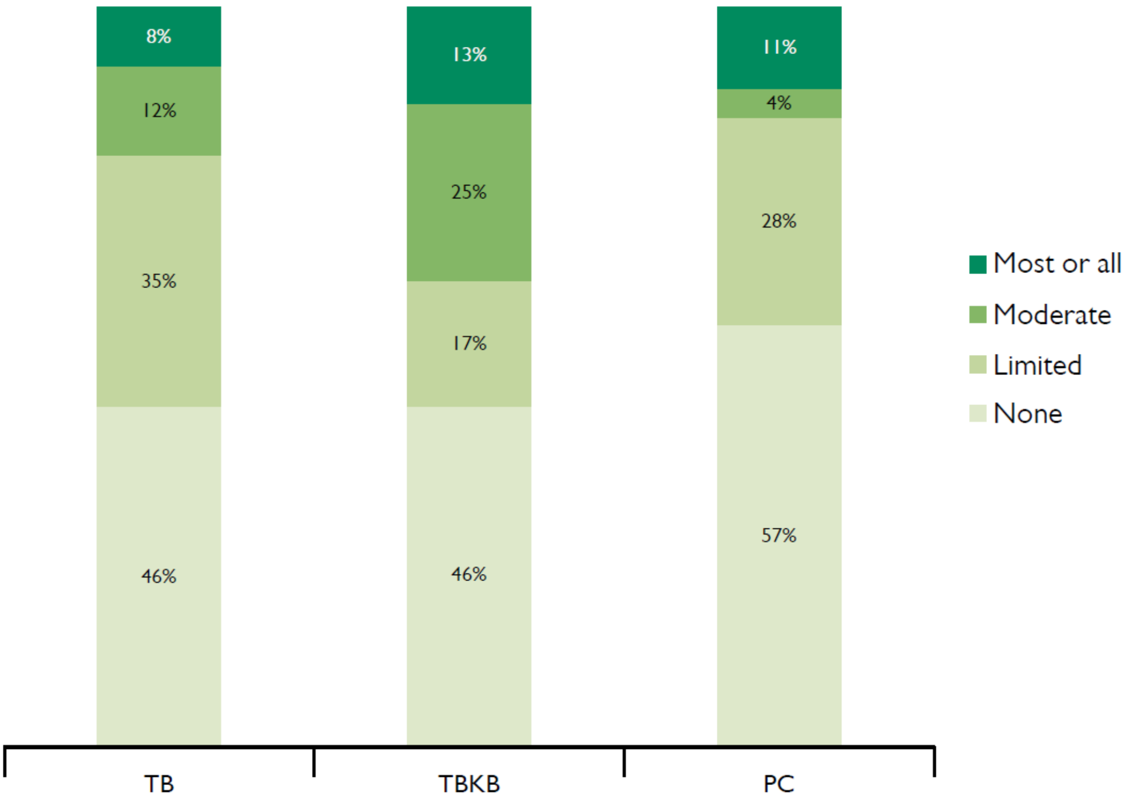
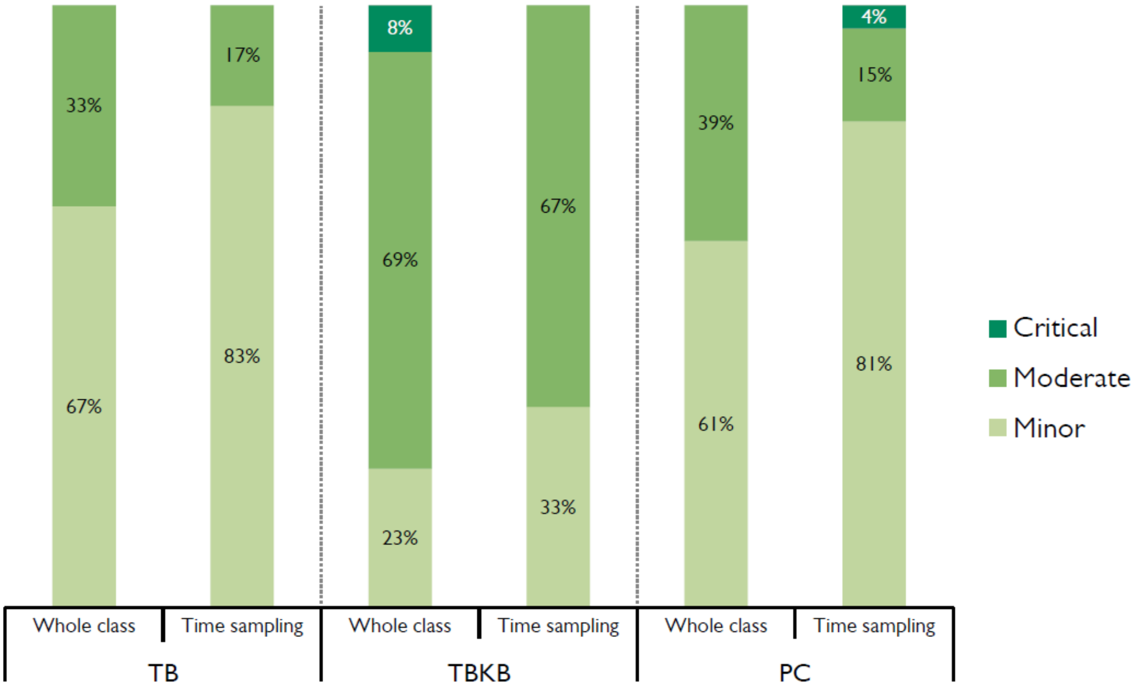


Figure 19
Severity ratings for reading passages *(Note: Percentages may not add to 100 due to rounding)*



Across all devices, most students disliked that they could no longer see the item after using the expansion tab to enlarge a passage. These students said that it was difficult to keep the item in their head while reading the passage and that they would prefer to read the passage and see the item at the same time. Preferably, the passage and item would sit side by side on the screen so that they could glance between the two when necessary. Additionally, some passages were initially presented to students in an expanded format. This meant that the first item relating to the passage was obscured and some students did not realise that they needed to hide the passage resource in order to get to the first item. Consequently, after reading the passage these students pressed 'Next' which took them to the second item. Once they became aware of the problem most students looked for the obscured item before advancing through 'Next'.

Students did not consistently use the expansion tab and the 'Show/Hide resource' button to expand and minimise the reading passage text. This was especially problematic for lengthier passages where the expansion controls were not always visible on the students' screens due to scrolling out of view. Students in the TB and TBKB conditions were frequently observed leaving the passage minimised in the left hand panel and using pinch and zoom to enlarge their view of the passage. Students in the PC condition were more likely to use the expansion tab to enlarge the passages.

The degree of readability for the reading passage had a direct relationship to the option students chose to enlarge the passage. Readability was improved when students used the expansion tab or 'Show/ Hide resource' button to expand the passage text to full size. Students who opted to zoom in on the passage text in the left hand panel had several challenges. Some Year 3 students made multiple attempts at adjusting the zoom which was often time consuming. Additionally, when TB users with low resolution screens enlarged a passage it became blurred. Lastly some students said they became confused when using the zoom, as it caused them to lose their place in the passage.

Some of the Year 7 and 9 students in the TB and TBKB conditions were observed 'swiping' across the screen to toggle between the expanded passage view and the item. This made it easier for these students to refer back and forth between the text and the question. While this strategy was most common with older students, some Year 3 students also used 'swipe' to switch back and forth between the passage and the item.

Across devices many students chose to read the passage text from the left hand panel without attempting to expand or enlarge it. This was despite the fact that they did not appear comfortable reading the very small font in some of the passages. Some Year 3 students who did not expand the text were observed answering the questions without having read the passage. When asked about this they said that the print was too small for them to read the passage properly.

Keyboards

Figures 20 and 21 show the frequency and severity ratings across all year levels for issues observed with the device keyboard. For the TB condition, this was the onscreen keyboard. For the TBKB and PC conditions this was the external keyboard. Issues with keyboard usage were observed somewhat more frequently in the TB and TBKB condition than in the PC condition. The severity rating for observed issues was highest for the TBKB condition, with more than half of the observed issues identified during the whole class observation indicated as 'moderate' or 'critical'.

Figure 20
Frequency ratings for keyboard usage *(Note: Percentages may not add to 100 due to rounding)*

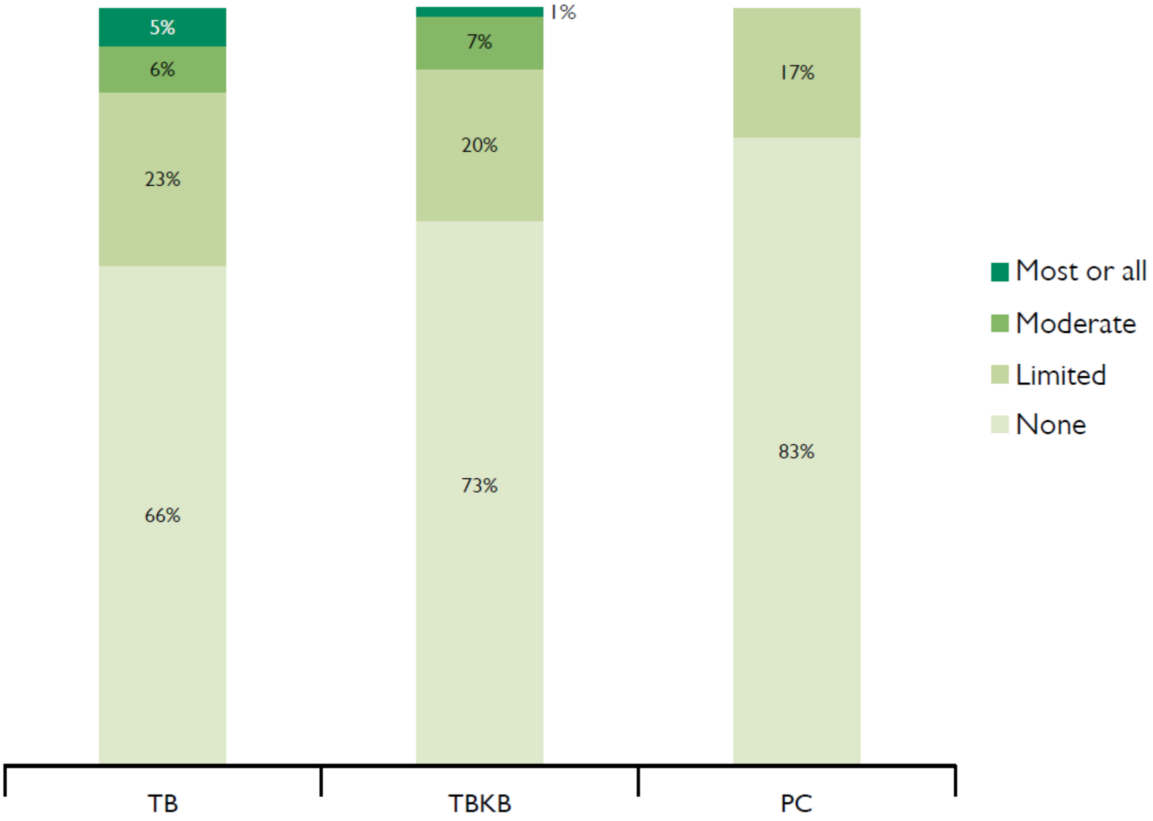
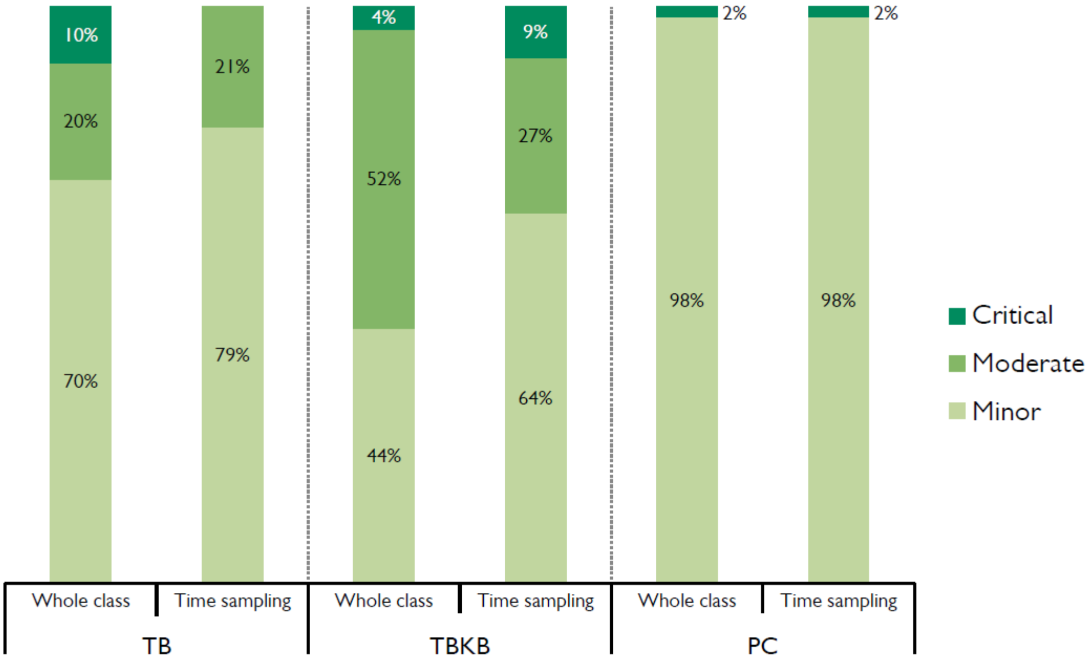


Figure 21
Severity ratings for keyboard usage *(Note: Percentages may not add to 100 due to rounding)*



Spelling

Figures 22 and 23 show the frequency and severity ratings across all year levels for issues observed with the audio functionality in the Spelling test. This includes challenges with playing or replaying the audio file, concerns with volume control, or excessive replay of the audio files by students. Interestingly, issues were more frequently observed with the PC condition than with the TB or TBKB conditions. The TB condition had the fewest issues observed, though some of these issues were observed for most or all students. The severity of issues was also greater for the TB condition with higher percentages of 'critical' issues observed than with the other device conditions.

Figure 22
Frequency ratings for audio functionality for spelling

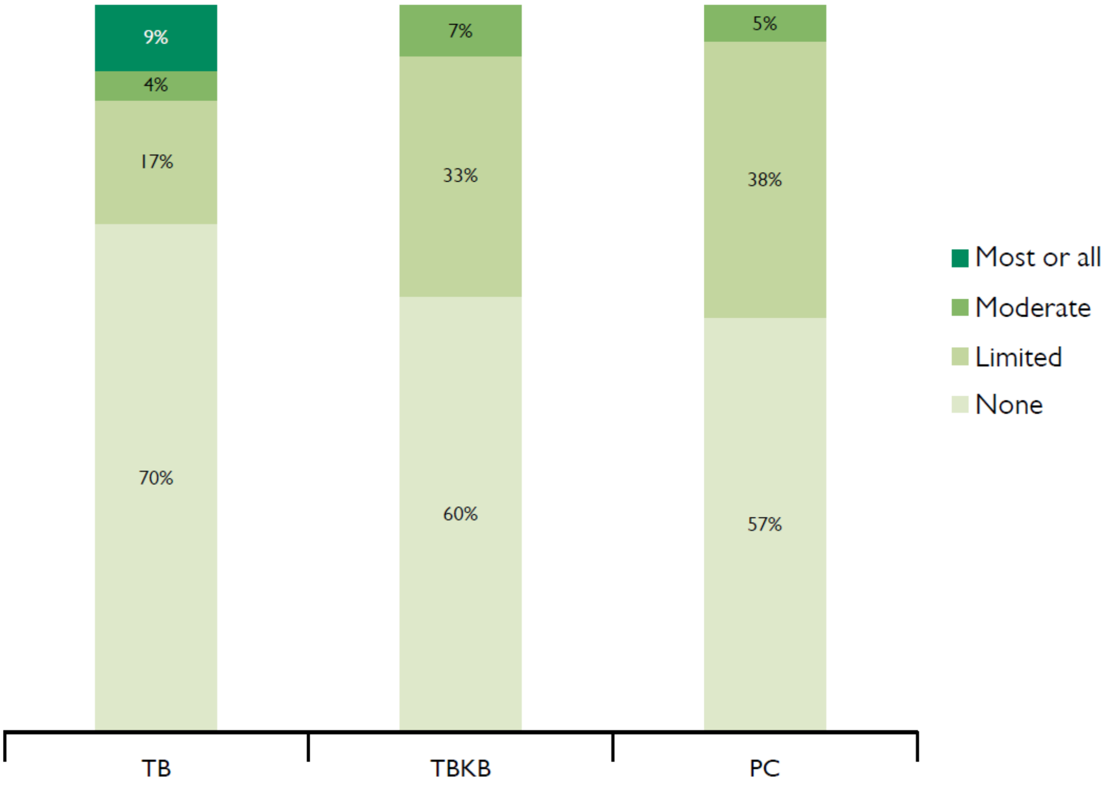
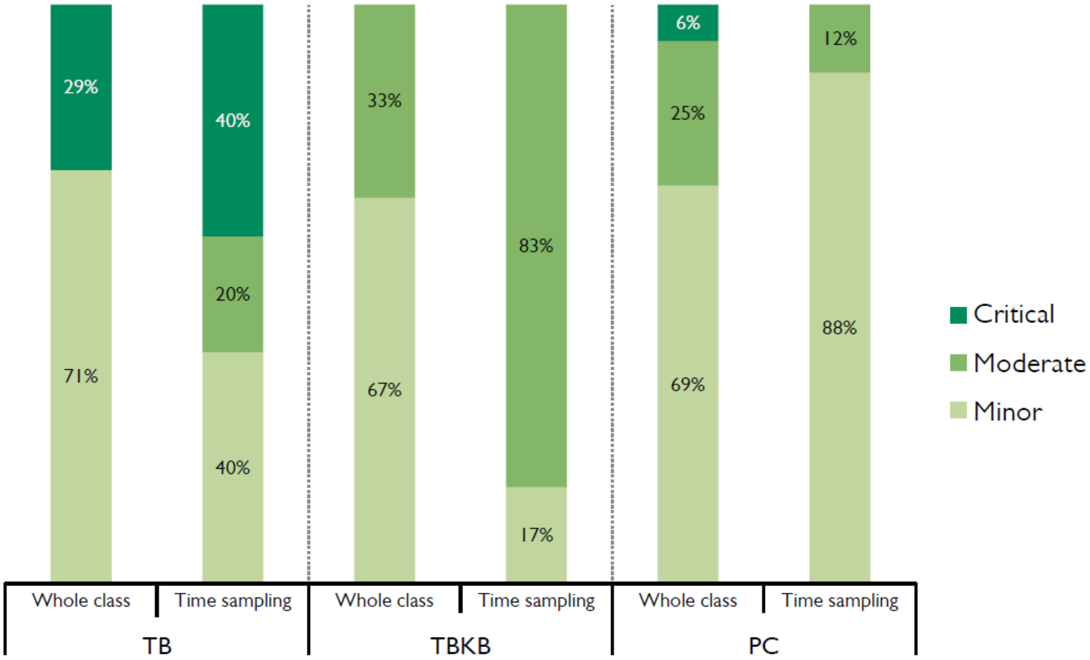


Figure 23
Severity ratings for audio functionality for spelling



Students generally said that they felt comfortable about doing the Spelling test online and could see that it would not be possible to do this type of test using pen and paper. Question 9 of the practice test illustrated how the audio function worked for the Spelling test. In encountering this item some students found that their device was on mute, or that their headphones were not connected properly, or simply did not work. It was beneficial to have had the opportunity to rectify this before the actual test.

On TB and TBKB devices the play button was too small for some users. These users needed to press multiple times to get their response to register. In some cases students' fingers obscured the play button. This was especially the case if TB devices were used in the portrait position. Otherwise, the audio function was described as easy to use. About a third of students said that they replayed the file when necessary and that they liked this feature. About 10% of students said that they adjusted the volume during the test. Most of these students were using tablets. Several students said that they had difficulties with audio clarity, noting that some words sounded louder or softer than others. They also mentioned a background crackling noise and that the highest volume setting was not loud enough. Some also felt that the voice mumbled at times and that the accent interfered with their understanding of the word. While students who raised these issues were able to complete the test, most said that they strained to hear the recording and often needed to replay it. Others said that they could not hear the word well enough and felt that they could be making errors as a result of this.

Headphones

Headphones were used during the Spelling test to listen to the audio. Across all devices, test administrators noted four main issues relating to headphone use. Firstly, some headphones did not work and had to be replaced3. Secondly, some classes had mechanical problems with the headphones such as connection in the wrong socket or loose connection in the socket. Thirdly, poor quality headphones affected some students' ability to hear the audio recording. Students adjusted to this by replaying the recording. The quality of headphones impacted on a few students' ability to hear clearly, but not enough for them to need to have them replaced.

Lastly, some headphones were too large and uncomfortable for some Year 3 students.

Results by item type

In each item type some of the items spanned more than one page and hence involved scrolling or zooming before students could read the item in its entirety. Students on each of the device types said this impacted their ability to respond effectively and efficiently. All item types required accurate positioning of the cursor, finger or stylus before a student could click, locate, drag and drop, or enter a response. The capacity to answer each item type depended on good screen visibility. A PC with a large screen, full sized keyboard and an external mouse was judged by students and observers as the easiest to use with all item types. However, this was sometimes only marginally easier than using a TB or TBKB (as with drag and drop items).

The practice test was intended to familiarise students with the functionality of various item type interactions. The content of the practice items, however, was typically very easy and as a result did not expose students to the full range of item interactions. Specifically, students claimed that they needed more information about drag and drop functions (e.g. actions needed to change answers), inline choice/drop down item functionality (e.g. the use of the arrow to select/change responses), and text entry items (e.g. how different responses would be marked relative to capitalisation, abbreviations, etc.). It is likely that a more robust tutorial and practice opportunity prior to assessment would have improved student interactions with item types across all devices.

Click item types

In general, students found click item interactions (multiple choice, multiple select, and hot spot) easy to use across all devices. For the most part, students displayed an adequate proficiency using the external mouse, touchpad, or with 'finger-as-pointer' to select responses to these items. The slight amount of extra effort required to select responses for TB and TBKB users did not seem to impact student performance on these item types.

Multiple choice items made up 75% of all Reading items and 25% of all Numeracy items on the study test forms. Students knew what to do with the familiar multiple choice item type. Although TB and TBKB users generally had little difficulty managing the touch screen interactions required for multiple choice items, some were noted having difficulty at times selecting the radio button precisely. Some Year 3 students accidentally triggered the zoom function when they tapped the screen twice. In some multiple choice items, not all of the options were visible on the screen at once which made answering these items harder for all device users.

Approximately 10% of items within the Reading and Numeracy tests were multiple select items. As with multiple choice items, most students knew what to do with the multiple select type items. Issues with student responses for this item type generally stemmed from failure to read directions closely (not realising they should have selected more than one response) rather than from user interface or device interactions. However, observers noted some issues with the scrolling function. This was especially the case where some answer choices may not have been visible on one page (perhaps exacerbating the issue of students not realising that they should select more than one response), and where the font size was too small to read easily.

Hot spot items made up a relatively small proportion (approximately 1%) of items on the Reading and Numeracy tests. All device users were able to respond easily to the hotspot items. The size of each of the 'hot' areas also made this easy for TB and TBKB users. Some students, however, said that they were unsure of how to deselect their initial response and choose another. Additionally, some hot spot interactions did not highlight or provide feedback that the response had been selected.

Inline Choice

Figures 24 and 25 show the frequency and severity ratings across all year levels for issues observed with inline choice items. Issues were more frequently observed in the TB and TBKB conditions than in the PC condition. In addition the severity ratings tended to be higher for the TB and TBKB conditions than for the PC condition.

Figure 24
Frequency ratings for inline choice items

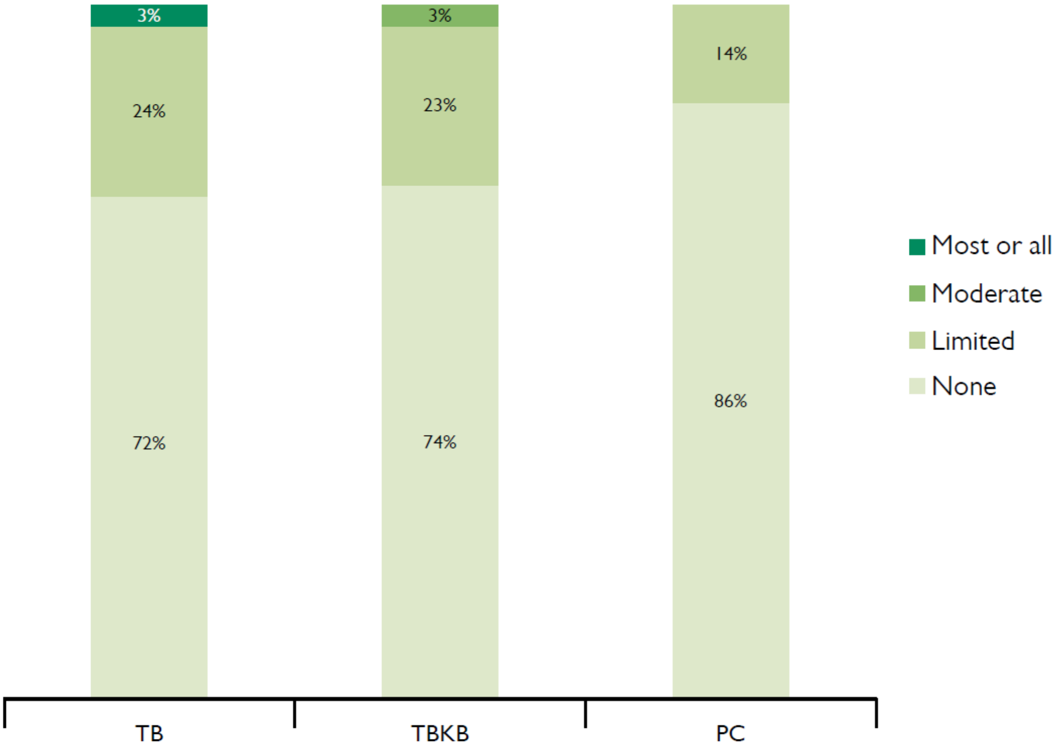
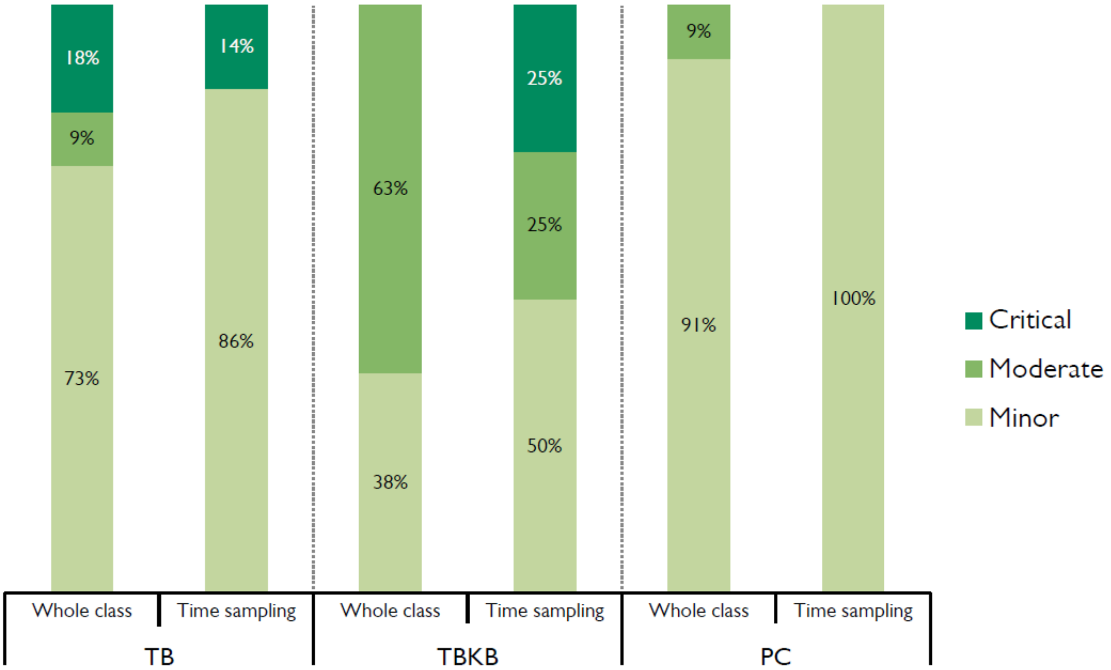
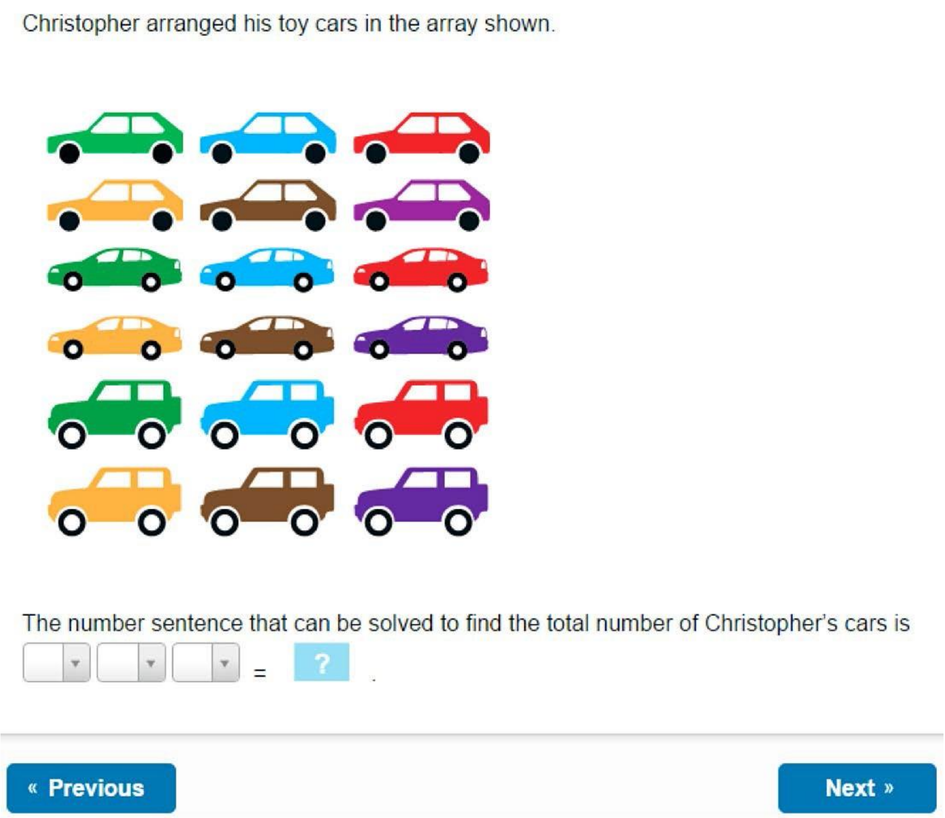


Figure 25
Severity ratings for inline choice items



Most PC users had little difficulty with inline choice items and some said that they preferred this item type to the drag and drop items. Test administrators noted that PC mouse users selected their choice more accurately from the drop down menu than did TB and TBKB users. TB and TBKB users had some challenges with the size of the inline choice/drop down menu field. These students reported using the pinch and zoom function to enlarge the menu field so they could read the options clearly and thus make an accurate choice. TB and TBKB users said that to make accurate selections they needed to spend time focusing carefully on the drop down menu. Additionally, the onscreen keyboard often popped up on TB and occasionally on some TBKB devices when students tried to open the choice menu. As a result some students attempted to type their answer into the open space rather than selecting a response from the menu. This observation, in particular, may explain why the Rasch analyses showed that the relative difficulty of inline choice items was greater for students in the TB condition than in the TBKB and PC conditions. It is not that the type of keyboard directly impacted student performance on this item type, but rather, that technical issues with the item interacted with the onscreen keyboard. Lastly, some TB and TBKB users accidentally hit the ‘Previous’ button as a consequence of its proximity to the inline choice interaction (see Figure 26).

Figure 26
Year 5 inline choice item


Christopher arranged his toy cars in the array shown.

The number sentence that can be solved to find the total number of Christopher's cars is

Drag and drop

Figures 27 and 28 show the frequency and severity ratings across all year levels for issues observed with drag and drop items. Across all devices, issues with drag and drop items were observed at a higher rate (in 46-62% of the observations) than with the overall frequency rating across behaviours (see Figure 11 earlier in the document). Additionally, issues were more frequently observed in the TB and TBKB conditions than in the PC conditions. The severity of issues tended to be somewhat greater in the TB condition than the PC condition and greatest in the TBKB condition.

Figure 27
Frequency ratings for drag and drop items

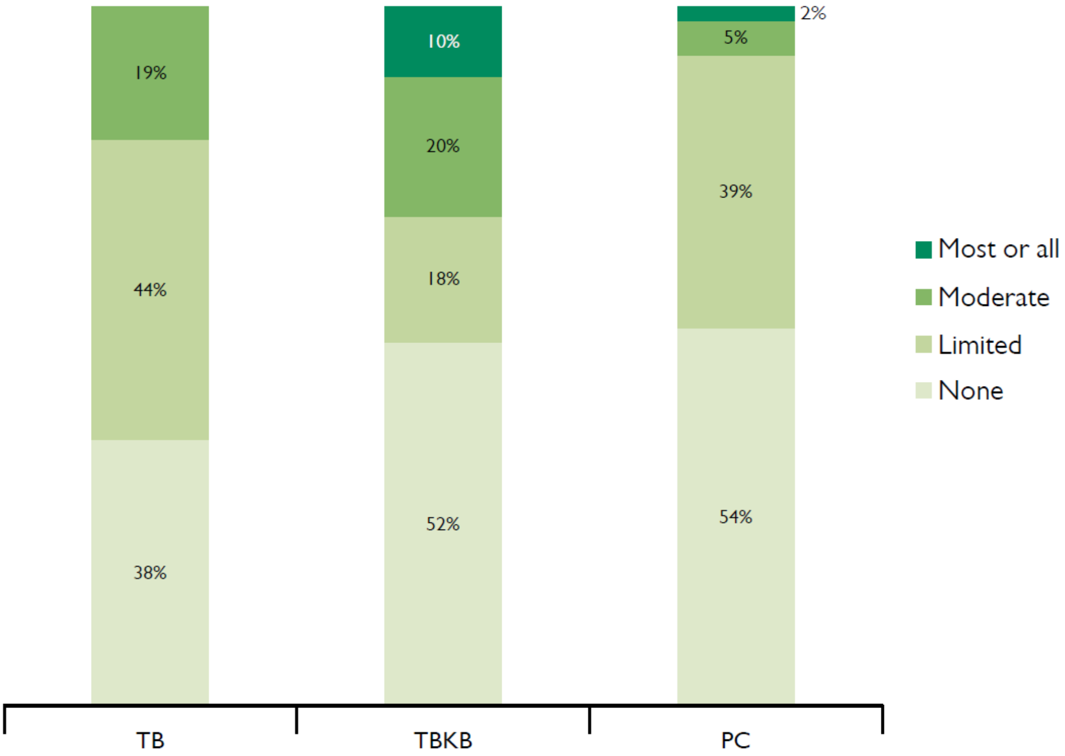
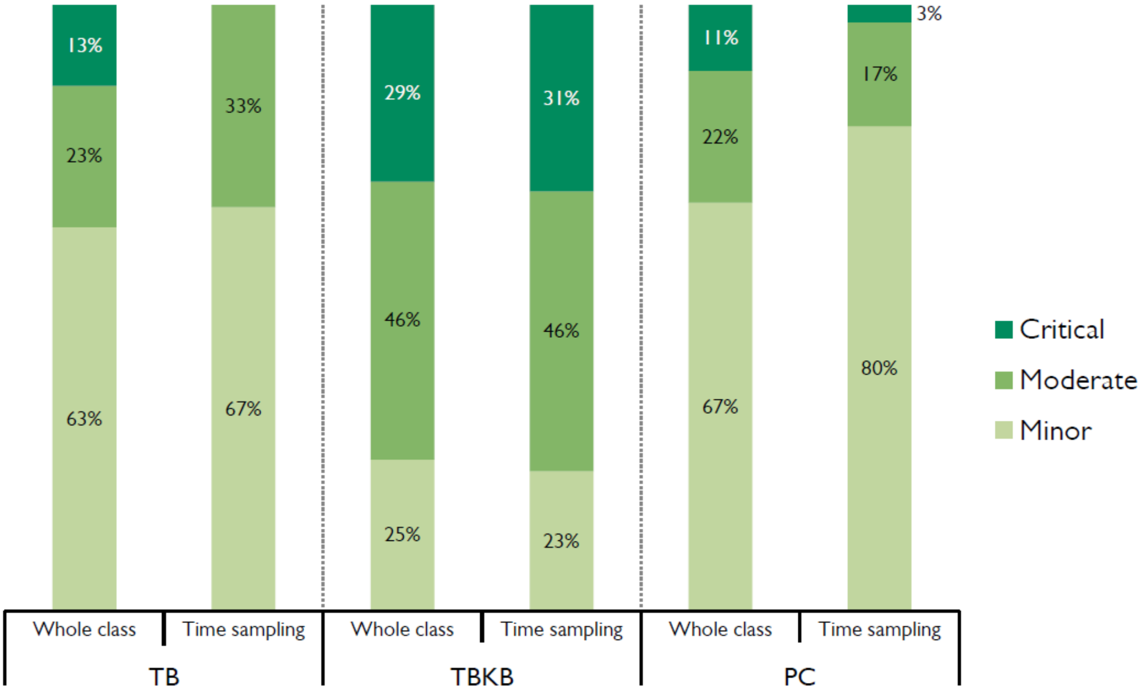


Figure 28
Severity ratings for drag and drop items

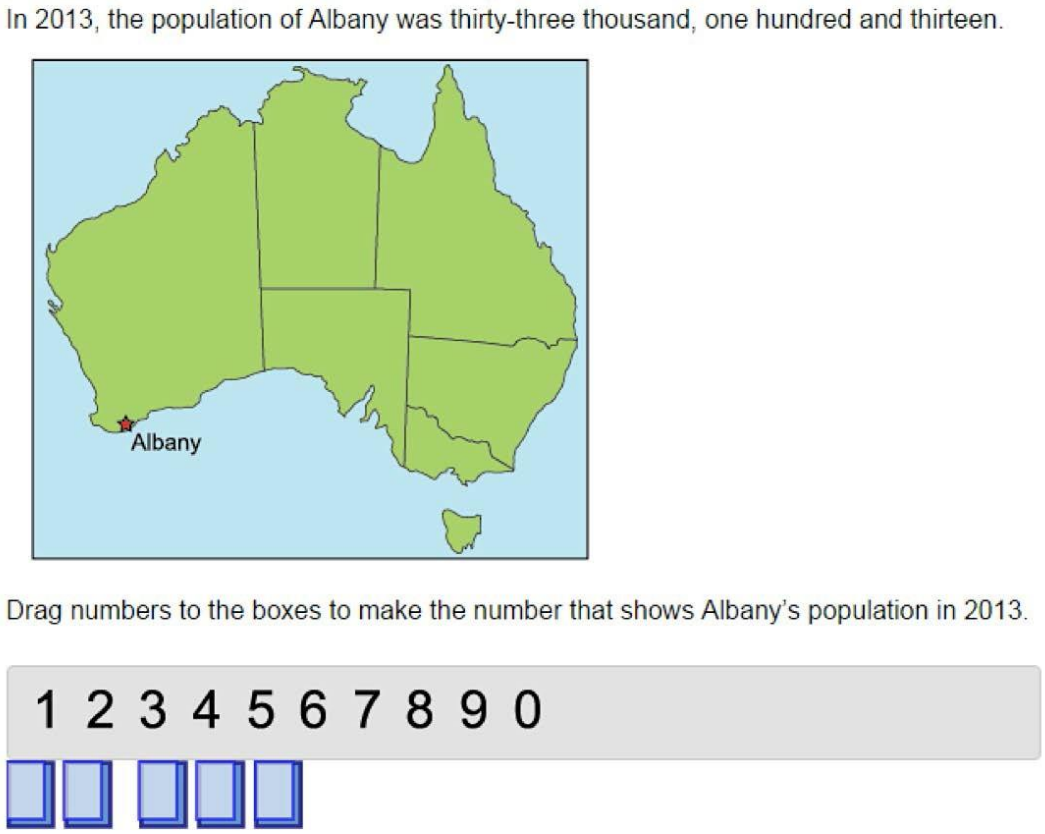


This item type was the most problematic of all the item types presented to students in this study. Across all devices, students asserted that they found the drag and drop functionality "frustrating, time consuming and difficult to use" and noted that this item type was "time consuming, irritating and required too many actions to achieve the goal". Additionally, many students were initially unaware of how to change their response. Some students said that it was confusing to have to return an incorrect response to the grey area before selecting the next choice. They also commented on the time required to change a choice once it had been dropped. Students suggested that an ‘undo last action’ button be installed to assist with changing responses. Several students said that it was ‘tedious and time consuming’ to enter more than four choices/entries. For example, Year 7 Numeracy item 17 required a small square to be dropped 15 times into the small drop zones of a column graph. Students suggested that the lower part of the column self-populate once the highest value is entered.

Although this item type was difficult for students across all devices, there were specific challenges with these items for students in the TB and TBKB conditions. Observers noted several possible reasons for this:

- Students’ fingers sometimes obscured the draggers and the drop zone.
- Students were unable to see the drop zone after zooming in on the dragger.
- Multiple attempts were required to drop the dragger accurately into the drop zone (e.g. TB users said that their choices would not drop properly into the pie graph).
- The draggers and/or drop zones were sometimes too small (see Figure 29).
- The draggers and/or drop zones were sometimes positioned too close together (see Figure 29).

Figure 29
Drag and drop item with size and spacing issues



In 2013, the population of Albany was thirty-three thousand, one hundred and thirteen.

Drag numbers to the boxes to make the number that shows Albany’s population in 2013.

Students in the TB and TBKB conditions found it very difficult to select small 'draggers' with their fingers and hold onto them while dragging them to the drop zone when their fingers obscured the dragger. In an effort to work around these issues, students would often zoom in to make the draggers and drop zones appear larger. However, students could not always release the dragger into the drop zone when they were zoomed in.

Students in the PC condition (especially students in Years 3 and 5) had challenges when an item required them to scroll between draggers and drop zone. Students using an external mouse had to hold the dragger by keeping the left mouse button pressed while using the mouse scrolling wheel. Students using a touch pad had to keep hold of the dragger by pressing on the touch pad while using the arrow keys to scroll. Although students generally managed to drag and drop their choices into the selected space, many said that it took multiple attempts to do so.

Text entry

Figures 30 and 31 show the frequency and severity ratings across all year levels for issues observed with text entry items. This includes short and extended response text entry items. Issues were observed somewhat more frequently for students in the TB condition than for students in the TBKB or PC conditions. However, the severity of issues was greater for students in the TBKB condition than in the TB or PC conditions.

Figure 30
Frequency ratings for text entry items

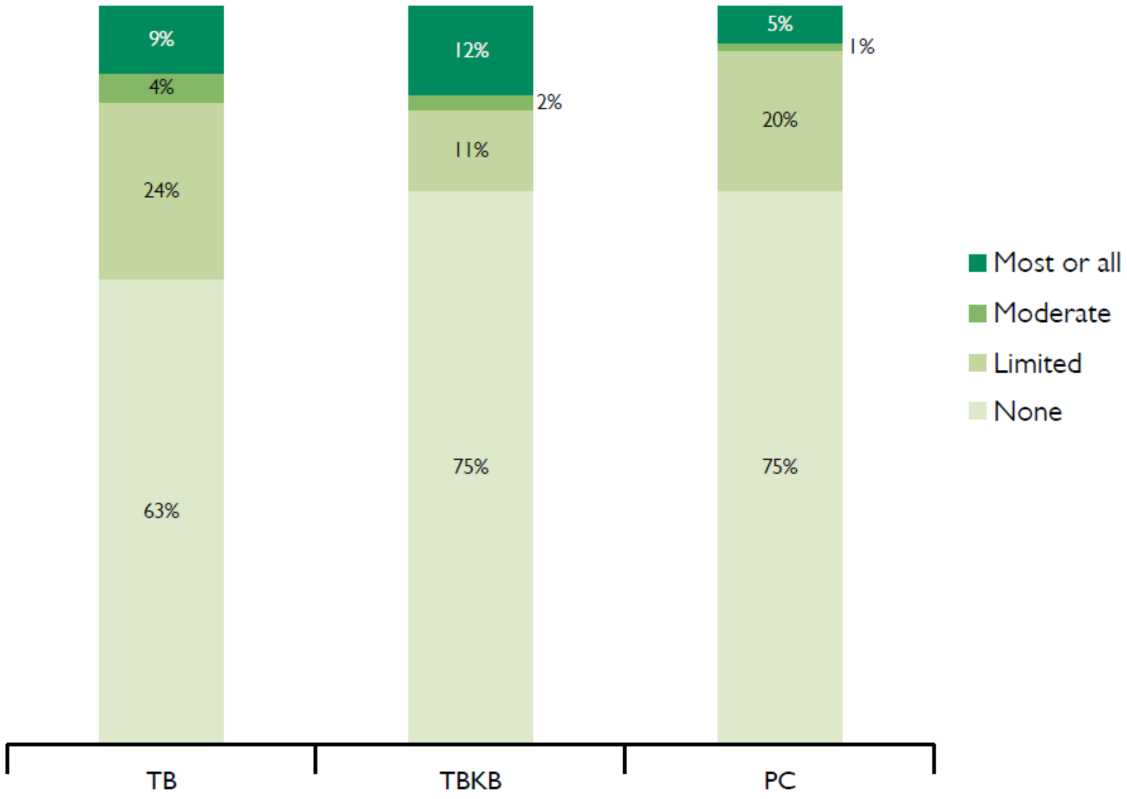
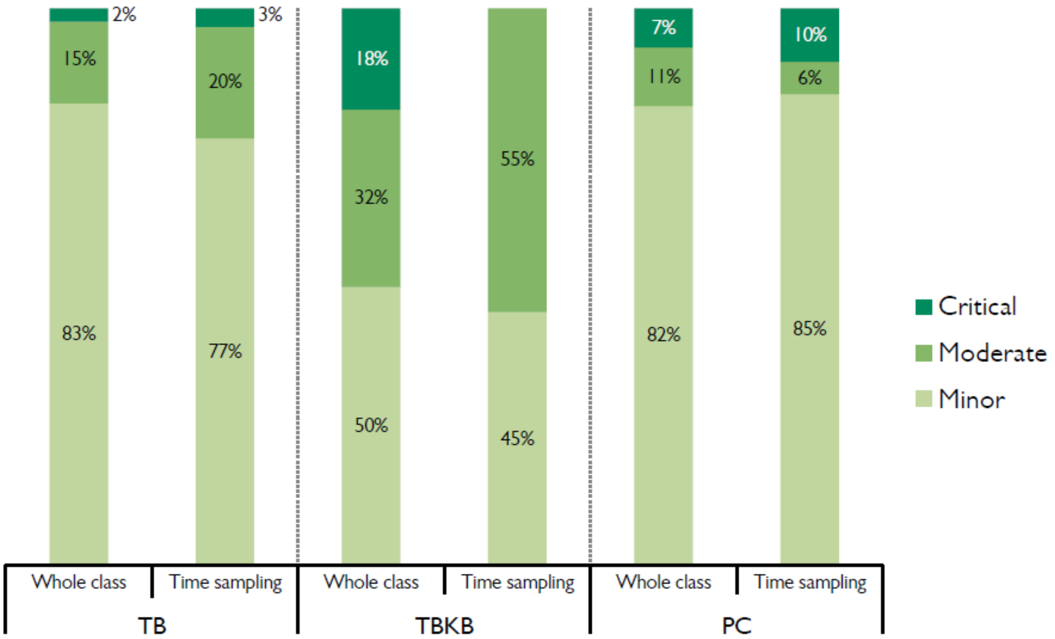


Figure 31
Severity ratings for text entry items



Across all devices, text entry for the short response spelling items was straightforward with no scrolling or zooming generally required. Students said that the text boxes were large enough to enter each word. For extended response items, students generally thought that there was enough space to write a response, however some students in the TB condition noted that the onscreen keyboard sometimes obscured part of this space.

Most students knew how to enter their response using the keyboard although some students did not initially know how to delete or edit their responses. Younger students were at a wide range of developmental stages in terms of their ability to use the keyboard to enter responses. Some Year 3 PC users said that they were comfortable using the full sized keyboards with large keys to enter their responses. Others were unfamiliar with the keyboard and had to locate keys one at a time. Some Year 5 students said they preferred to type rather than hand-write their response, while others said they lacked keyboard skills. Proficient keyboard skills underpinned Year 9 students' ability to deal this item type.

General device interactions

PCs

Students who used an external mouse during the tests said that it was particularly helpful with drag and drop items, 'click responses' and navigation. Touchpads proved a challenge for students especially when they were unfamiliar with them. Students who were less familiar with touchpads reported that they would lose the cursor or would not be able to get the cursor to go where they wanted. Additionally, novice touchpad users said they had difficulty using it to scroll the page up to be able to see the available options. However, students who were more proficient in using the touchpad did not have the same issues with accuracy and speed. Furthermore, students who used both hands on the touchpad were much more efficient than those using one finger.

Tablets

Consistent with previous research, most students seemed to prefer using their TB or TBKB in landscape orientation. However, landscape orientation often resulted in 'whole-of-page' visibility issues and some students rotated their tablets to portrait orientation to improve visibility. This created other issues such as smaller font size and an increased need to zoom to enlarge the text and other item features. In some cases, once students had rotated their tablet to portrait orientation they kept it that way for the remainder of the testing session. In other cases, students would switch between landscape and portrait positions depending upon the specific needs of each item. Students in the TBKB condition kept their tablets in landscape orientation more often than students in the TB condition due to restrictions presented by the tablet stand or case.

TB and TBKB students were observed using their devices in a variety of positions. Some TB users used their device on its stand. Others generally used the device while it was flat on the desk or rested it on the desk while holding the device in their hands. To provide a slight angle, tablets were observed propped up within, or on, the device's carrying case. Those who preferred a more vertical angle used their magnetic flip screen cover as a stand. Some TB users worked with their tablets on their laps. In a small number of cases students held the tablet up with one hand and used the other hand to interact with the items (students noted that this was easier if the device was in portrait orientation). The TBKB condition offered somewhat fewer device positioning options due to the presence of the external keyboard and the tablet stand or case. However, during the Reading and Numeracy tests some TBKB users either removed their detachable keyboard, or disengaged their wireless keyboard in order to use the tablet on their laps. The variety in device positioning observed in this study was somewhat greater than what had been observed in previous research.

Students did not always keep the same device positioning throughout the test. TB users at all levels changed the position of their device throughout the testing period to suit personal preferences and the requirements of each test. By contrast, TBKB users across all levels and tests made fewer changes to the position of their device throughout the testing period than TB users.

Students in the TB and TBKB conditions had mixed feelings about using their fingers to interact with the test content. While some students said that it was easy, others thought this made it difficult to select elements accurately and noted that their finger sometimes obscured key parts of the screen. Additionally, students (especially younger students) were observed to trigger functions accidentally in the text due to spurious touches or imprecise selection. Students in the TBKB condition had slightly more challenges using their fingers to interact with the test content than students in the TB condition. This may have been due to the requirement to transition between the touchscreen and the external keyboard.

In contrast to previous research, students in this study made fairly heavy use of pinch and zoom functionality across most domain areas to help them manage the visibility of test content. This was especially the case when content involved panels which had to be expanded or minimised or when scrolling was required to see all parts of the item. Pinch and zoom was often preferred to using the controls built into the user interface. Students did not report using pinch and zoom within the Spelling domain. This is likely to be due to the fact that there were significantly fewer issues with visibility of the test content in the Spelling domain.

While pinch and zoom provided a popular mechanism for students at all year levels to manage the test content on screen, the degree of success and challenge students had with this feature varied. Younger students generally made greater use of the pinch and zoom feature than did older students. They were, however, more often observed having difficulty controlling the zoom than the older students. Specifically, some of the younger students said that they sometimes had difficulty controlling the zoom feature as it would 'zoom out' when the screen was tapped twice. Additionally, some students at all year levels said that it was difficult to relocate to the question if they zoomed out too far. Lastly, image quality was an issue when using the pinch and zoom feature, in that some images became fuzzy and unclear when enlarged.

Ergonomics

Feedback from PC users at all levels and in all domains indicated fewer problems relating to fatigue or strain from device usage than with TB and TBKB users. Specifically, PC users preferred looking at the screen at eye level rather than having to look down at a screen. These students reported that being able to use an external keyboard and mouse was generally comfortable and did not result in physical strain.

The most common problem reported by tablet users was the small font. The need to squint, problems with eye strain, and concerns about small and fuzzy font (particularly relating to the reading stimulus) were reported at all year levels. The use of a smaller screen increased the incidence of students reading very close to the screen (especially for Year 3 students). As the amount of text increased across year levels, the issue with small font size become more pronounced. Students in Years 7 and 9, however, were more proficient in their ability to use device features to overcome some of these problems than younger students.

A small number of students in the TB and TBKB conditions complained about screen glare, while others were observed using low levels of screen brightness. Some Year 3 students reported that their chairs were uncomfortable or were set at the wrong height. This could indicate a mismatch between the physical configuration of the classroom space for daily activities and the ergonomic needs of specific device use for extended periods of time. Some students indicated that they stretched to relieve muscle tiredness. Others noted that they had been taught to stretch regularly while working with electronic devices in order to overcome tiredness.

Year Level

The range of computing skills within each year level varied considerably. At all year levels there were students who were confident device users and prepared to problem-solve any issue they encountered. However, Year 3 students stood out from the other year levels in regard to the additional time required for them to understand the unfamiliar device features. Compared with students in the other year levels, Year 3 students generally:

- needed more time to become familiar with their device and its specific features, the testing interface and how to navigate the testing program (frequently asking the test administrator for assistance),
- had more difficulties positioning the cursor on a touchscreen device and would have liked the option of using a stylus,
- were slower to enter text due to their lack of keyboard familiarity and skills,
- more frequently lost control of the zoom function and had difficulty finding their place on the page,
- more frequently triggered the zoom or other functions with their finger accidentally,
- were more surprised by, and less able to deal with, any unexpected technical or device related issues when they arose; and
- more frequently used invalid actions on their device such as trying to swipe a passage to close it, or using the ‘Enter’ button to try and advance to the next item.

At other year levels, where students encountered similar issues relating to devices, they were better equipped to deal with them due to greater proficiency with electronic devices and digital interfaces in general.

Device Familiarity

Participating classrooms were categorised as to whether or not they had prior experience using TB, TBKB, or PC in classroom activities. Observer frequency and severity ratings were compared for those instances where the testing device matched the device used in the classroom and those where it did not. Overall results were very similar between classes familiar with and unfamiliar with the testing device. There was, however, some reduction in the observation of issue frequency and severity for classes familiar with the device in areas identified as challenging for students. This suggests that increased exposure to the testing device could ameliorate many of the issues observed.

Specifically, classes familiar with the testing device tended to have fewer observed issues with:

- general navigation,
- reading passages,
- onscreen keyboard (for TB); and
- drag-and-drop and text entry items.

Additionally, classes familiar with the testing device tended to have less severe issues with:

- general navigation,
- reading passages,
- external keyboard,
- drag-and-drop and text entry items,
- using the finger as a pointer; and
- tablet stands/cases (for TBKB).

Interestingly, classes familiar with the tablet with an external keyboard tended to have more issues with the onscreen keyboard than those who were not familiar with the external keyboard. This suggests that once students become familiar with the external keyboard, it may be preferable to the onscreen keyboard.

5. Conclusions

5.1 Summary of findings

Overall results from both psychometric and qualitative analyses support the conclusion that device effects, when present, are not pervasive but are centered on specific item types, interactions, or interfaces and affected by device familiarity or experience level with online testing. The salient factors relative to the observation of device effects appear to be:

- use of and familiarity with an external keyboard (especially for older students),
- amount of scrolling required by the test domain interfaces (especially for Reading),
- specific characteristics of drag and drop items; and
- general familiarity with the device, test interface, and item interactions.

The Rasch analyses generally found no difference in performance across device conditions for Years 3 and 5 for Reading and Numeracy. There were differences observed when comparing PC to TB for Years 7 and 9 (favouring PC) but these differences were largely mitigated through the use of an external keyboard, such that performance in the PC and TBKB conditions was generally not significantly different. These results seem to be driven by text entry items where the external keyboard appeared to have a positive impact on performance. The GLMM analyses additionally suggest that students who were already familiar with external keyboards were potentially disadvantaged by not having access to one during testing.

The frequency and severity ratings from the observation data were often higher for the TBKB condition than for the TB condition which, in contrast to the psychometric results, suggests potential concerns with the use of external keyboards. However, test administrators reported that external keyboards had compatibility issues with some of the tablets that may have made certain tablet functions (e.g. touch screen interactions, tablet rotation, etc.) more challenging. Additionally, the test interface was not optimised for tablet use and the complexity of interactions required to navigate within the interface while at the same time interacting with the external keyboard may have had a negative impact on student experiences. Both of these factors combine to make it difficult to disentangle the effect of using an external keyboard from other concomitant variables which influenced the usability and utility of the external keyboard. If schools had more opportunity to test external keyboards with their tablets, students were provided with more experience using the external keyboard, and the test interface was optimised for tablet interactions, observations might have yielded different outcomes.

Another salient factor relevant to the observation of device effects in this study was the amount of scrolling required by the test interface in each domain. In general, the amount of scrolling required to see each item in its entirety as well as to be able to access the navigation controls (within an item as well as between items) was a concern across all domains except for Spelling, where scrolling was generally not observed so long as the devices met technology requirements for screen size. Vertical scrolling was present in Numeracy due, in large part, to the size of the items and the need to display graphic materials (e.g. tables, charts, etc.) along with item stems and response options. In addition, Numeracy tests reflected the greatest number and widest variety of Technology Enhanced Items (TEIs) of any domain in the study and, therefore, tended to have complex response formats which consumed more space.

While Reading did not have the same ‘item size’ issues observed in Numeracy, scrolling issues were likely exacerbated by the test interface for displaying reading passage text. Specifically, the length of reading passages introduced scrolling effects that made it difficult for some students to toggle back and forth between the question/response area in the right-hand panel and the stimulus material (reading passage) in the left-hand panel. As a result, students did not consistently use the intended controls (expansion tab and the ‘Show/ Hide resource’ button) to expand and minimise the stimulus materials. Instead, students in the TB and TBKB conditions were frequently observed leaving the left hand panel minimised and using pinch and zoom to enlarge their view of the stimulus material. Alternatively, students who did use the intended controls to expand/minimise the stimulus materials frequently had to rotate their tablets from landscape to portrait orientation or scroll extensively to be able to access them.

The frequency of use of pinch-and-zoom and screen rotation was observed at a higher rate in this study than had been observed in previous research (Davis, 2013; Davis, Strain-Seymour, & Gay, 2013). This further highlights that the test interfaces may not have been ideal for use with tablets. If, however, the test interface were improved and optimised across devices, the need for these actions could be significantly reduced. This might further improve comparability of student experiences across devices by reducing the frustration and effort involved in these additional steps.

The Spelling test generally offered the most usable and efficient interface for students of all the domains included in the study. Issues related to headset and audio configuration were, for the most part, well managed and did not impact the student testing experience. Some students were challenged by the clarity and volume of the audio recording, but most were able to manage this with the controls provided.

Psychometric analyses consistently revealed differences in performance across devices for drag and drop items in both the Reading and Numeracy domains. The device effect for the drag and drop items appeared to favour students testing on TB in earlier Year levels (3 and 5) and students testing on PC or TBKB in later Year levels (7 and 9)—though it is unclear as to why performance on drag and drop items would have been improved by the addition of an external keyboard. The observation data confirmed that the drag and drop item type presented the most difficulties for students of any item type included in the study.

This was particularly true for TB and TBKB users and for non-proficient touchpad users. These issues were encountered when the draggers and/or drop zones were small, when large numbers of dragging actions were required, and when the drop zone was out of sight or too close to the navigation buttons.

Student familiarity with the device used in testing was another salient issue and reduced the frequency and severity of issues observed in many of the key areas where challenges were observed. Test administrators observed fewer issues when the methods for text input (onscreen or external keyboard) or object selection/manipulation (finger, mouse, or touchpad) was familiar to students. In addition, familiarity with the specific test interface and item interactions is also important. While the practice test was intended to familiarise students with the overall test interface and functionality of various item type interactions, the practice test in this study was quite brief (only 9 questions) and did not give students a sufficient opportunity to become comfortable with the test interface and item interactions. It is likely that a more robust tutorial and practice opportunity prior to assessment would have improved student interactions with item types across all devices.

The lack of differences in performance across devices for students at Years 3 and 5 might seem surprising given that younger students have less experience with testing, in general, have less well developed motor skills which can impact their ability to precisely control the testing interface, and generally lack experience with text entry. Certainly, these findings stand in contrast to the observational data for Year 3 students in the current study as well as to prior research (Strain-Seymour and Davis, 2013; Davis, 2013) which both show that younger students are more likely to encounter usability issues, less able to recover from these issues, and less likely to persist and try different approaches than older students. However, younger students are also likely to have grown up in environments where digital devices of all types are ubiquitous. An 8-year old Year 3 student has not lived in a world without the iPhone. Certainly many parents have had the experience of their own children being more facile with a device than they are. Perhaps this saturation of technology throughout early childhood development has supported a plasticity or resilience in younger students that helps them to overcome barriers even while simultaneously being more aware of them.

With specific regard to text entry, digital text entry on any device tends to be a less familiar task for younger students than for older students. This makes text entry items challenging for younger students regardless of device and, consequently, makes it less likely that differences across devices would be observed.

ACARA is confident that the information provided in this study will allow construction of a NAPLAN online assessment that accounts for any incidental overall device effects assuming student familiarity with the online platform and the device used for the assessment.

5.2 Recommendations

Within this section, recommendations are made that are intended to support successful student interactions across a range of devices by addressing the salient factors identified in the report summary.

- Review and revise testing interfaces to minimize the amount of scrolling needed to interact with test content and navigation controls within an item.

  Be mindful that it is not only the size of the item itself which can generate scrolling, but also the length of reading passages and response area for extended response text entry. Having a testing interface which supports ready and accessible navigation between the components of an item is an essential foundation to supporting device comparability.

- Conduct usability research with updated testing interfaces using PC, TB, and TBKB device configurations.

  Research should evaluate usability of features, in general, as well as identify issues which might be device specific. Usability testing which includes only one device configuration is likely to overlook device specific issues.

- Re-evaluate and revise drag and drop item interactions with specific focus on touch-screen tablets and touchpads on laptop PCs.

  Draggers and/or drop zones should be sufficiently large to accommodate the lesser degree of precision and visual obstruction presented when students use their fingers as a pointer rather than a mouse. When possible, effort should be made to ensure that drop zones and draggers are visible to students without scrolling and are sufficiently removed from the navigation controls to avoid accidental triggering of these controls. Interactions that require a large number of draggers to be repetitively moved should be re-evaluated to determine whether a different type of interaction would be better suited to assessing the content standard.

- Conduct additional research to study the impact of the external keyboard under conditions which incorporate updated testing interfaces and which mitigate technical issues related to keyboard compatibility with tablet devices and control for student experience with external keyboards.

  This research might focus on use with older students as their familiarity with text entry and keyboarding skills makes them most likely to receive benefit from the external keyboard. The study design might explicitly consider the length of the student response as a variable to be manipulated by including items that might elicit different response lengths and by providing students with motivation to attempt lengthier responses.

  Develop tutorials and practice tests which incorporate updated testing interfaces and which can be made available to students and schools in advance of testing.

  These tools should include a full range of item types and interfaces students will encounter. Tutorials might be developed with simplistic items to illustrate the basic functionality of the testing interface, but practice tests should include a realistic representation of test content and difficulty so that students will be encouraged to interact with the test content in a realistic manner.

- Recommend to schools that an external mouse be provided for students testing on laptop PCs.

  External mice are usually the only pointer device option available for desktop PCs. However, this equipment may be easily overlooked with laptop PCs as these devices usually include a touchpad or other built-in pointer device. The touchpad provided a number of challenges for students in this study especially when they were unfamiliar with its use. Providing students using laptops with an external mouse would give them the option of using either the external mouse or the touchpad as per their preference. External mice are relatively inexpensive and provide a high degree of precision for selecting and manipulating objects within the testing environment

5.3 Study limitations

As is typical with research studies conducted in school environments, the actual conditions surrounding data collection are not always consistent with what was planned. These differences are important to note to help provide context for interpreting the study results.

Technical issues

A range of technical issues were encountered in most sites during the study which impacted students' abilities to login to the test, work with the testing device, and interact with and respond to the items. These issues reflected both hardware and software challenges. It was sometimes difficult for observers to determine whether observed issues were related to the use of the device itself or were the result of technical problems. Hence, the test administrators' ability to rate the frequency and severity of some device related issues accurately may have been influenced by these perceptions.

The user interface

The testing software used in this study employed a set of user interfaces for each domain which did not reflect the final set of user interfaces that ACARA intends to implement operationally. Many of the test administrators' observations about student device use were as a result of students trying to navigate these interfaces on devices for which they had not been optimised. The frequency of use of pinch-and-zoom for TB and TBKB users, the amount of scrolling required for all devices, and the frequency of rotation of devices from landscape to portrait are all likely to be the result of the specific interfaces made available for the study. It is likely that these behaviours would be observed with much less frequency with a more finalised set of user interfaces.

The Spelling test

The Spelling test was shorter than the other testing domains having just nine items. Therefore, even with some initial set-up time, students still completed this test very quickly. The shorter duration of this test had the effect of limiting the opportunity for test administrators to observe student interactions. Additionally, the spell check and auto correct functions were often turned on and hence the performance data relating to the Spelling domain were compromised. As such, the amount of information available to evaluate any device effect for the Spelling test is more limited than in the other domains.

Device classifications

There were several issues which affected the interpretation of study results relative to device condition. Firstly, there were some devices which did not fit clearly into one of the pre-determined device classifications (PC, TB, or TBKB). These devices tended to be convertible PC/tablets which had a touch-screen interface but could also be used as a laptop with integrated external keyboard. For analysis purposes, these devices were grouped into the TBKB group but differ from actual tablets with external keyboards in both size and physical structure.

Secondly, there were some schools that, on the day of testing, did not have the number of devices expected. Consequently, some students in these classes took the tests on devices other than those planned. This resulted in several ‘hybrid’ classroom observations where test administrators were observing students working on different devices in the same classroom. For example, a classroom of 25 students might have had 10 students working on PC and 15 working on TB. Test administrators differentiated between devices in the hybrid classrooms as much as possible, however this might have impacted their observations.

Lastly, although schools were asked to provide tablets with a minimum 10 inch / 254mm screen size, this was not always the case. In some cases tablets with smaller screen sizes, such as the iPad Mini, were used by students during the study. The small screen on the iPad Mini was considered a limiting factor in terms of visibility and users’ ability to navigate effectively and to enter their responses accurately.

Space configuration and test security

Some desk configurations were more conducive to the testing process than others. In some cases students were able to see other students’ screens clearly. At times throughout the study, when students were unable to execute a function on their device, they automatically sought help from their neighbour as might be their normal classroom practice. In these contexts they were able to see other students’ answers. Despite this issue, cheating was not evident. In fact, the opposite seemed to be the case, with students being very concerned about potentially submitting work which was technically or otherwise assisted. In general, students demonstrated a concern about security and wanted the issue above to be rectified in ‘real test’ contexts.

Student motivation

In any research context the level of effort provided by participants can influence the results. For the most part, students in the study seemed to take the testing experience relatively seriously. However, on the extended response items in the Reading domain, some Year 7 and 9 students were observed cutting a section from the passage and pasting it in as their extended response. While there is no evidence to suggest that students were more motivated in one device condition than another, evaluation of both quantitative and qualitative results should, nonetheless, take this into consideration.

6. References

Bennett, R.E. (2003). *Online Assessment and the comparability of score meaning* (ETS-RM-03-05). Princeton, NJ: Educational Testing Service.

Davis, L.L. (2013). *Digital Devices February 2013 Usability Study Report*. Unpublished technical report, Pearson.

Davis, L.L., Orr, A., Kong, X., & Lin, C. (2015). *Assessing student writing on tablets. Educational Assessment, 20*, 180-198.

Davis, L, L. Kong, X. & McBride, Y. (2015). *Device comparability of tablets and computers for assessment purposes*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Davis, L.L., Strain-Seymour, E., & Gay, H. (2013). *Testing on tablets: Part II of a series of usability studies on the use of tablets for K-12 assessment programs*. Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/Testing-on-Tablets-Part-II_formatted.pdf

Findlater, L. & Wobbrock, J. O. (2012). Plastic to Pixels: In search of touch-typing touchscreen keyboards. *Interactions, 19*(3), 44-49.

Keng, L., Davis, L.L, McBride, Y. & Glaze, R. (2015). *PARCC spring 2014 digital devices comparability research study*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Olsen, J.B. (2014, April). *Score comparability for web and iPad delivered adaptive tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.

Strain-Seymour, E., Craft, J., Davis, L.L, & Elbom, J. (2013). *Testing on tablets: Part I of a series of usability studies on the use of tablets for K-12 assessment programs*. Retrieved from http://researchnetwork.pearson. com/wp-content/uploads/Testing-on-Tablets-PartI.pdf.

Strain-Seymour, E. & Davis, L.L. (2013). *PARCC device comparability, part I: A qualitative analysis of item types on tablets and computers*.

Way, W.D., Davis, L.L., Keng, L., & Strain-Seymour, E. (2015). From standardisation to personalisation: The comparability of scores based on different testing conditions, modes, and devices. In *Technology in Testing: Measurement Issues*, ed. F. Drasgow. Vol 2 of the NCME book series.

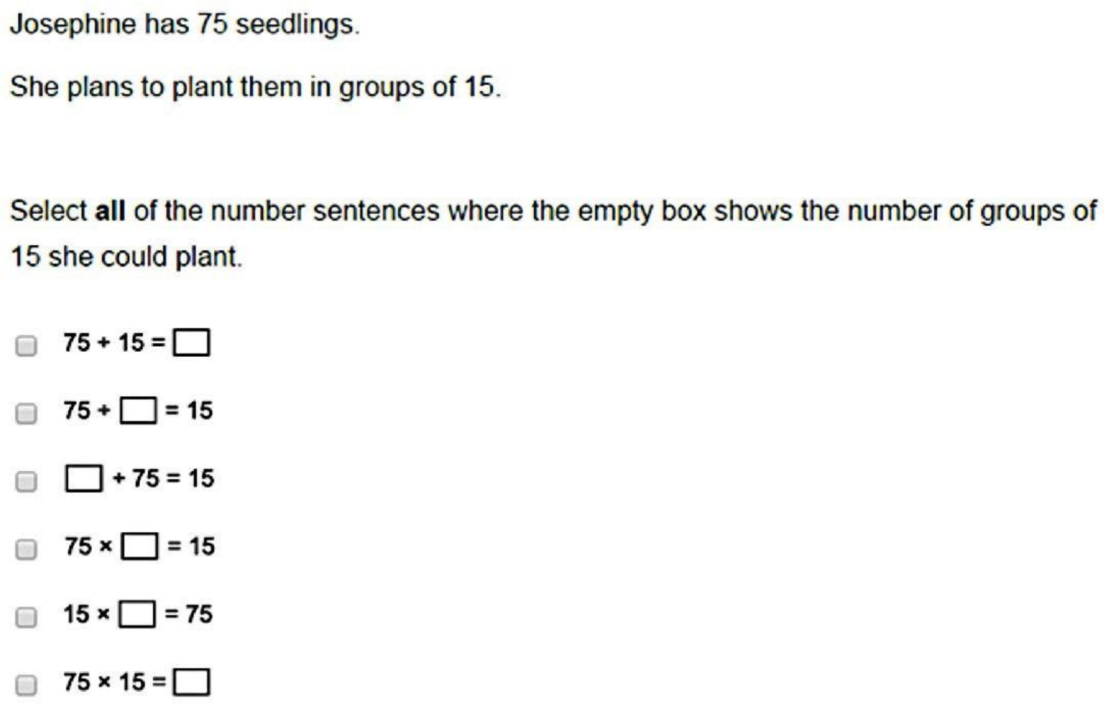
Wu M.L, Adams R., Wilson M.R., Haldane, S.A. (2007). *ACER ConQuest: Generalized item response modeling software (version 2)*.

Appendix A
Definition of item types

Item type	Item description / definition within an online context
Multiple choice	An item response type which involves clicking one of the radio buttons associated with the range of options provided. Multiple Choice items generally contain 4 options. Once an option is chosen, its button remains highlighted and any changes result in the last button clicked being highlighted and registered.
Multiple select	An item response type which involves clicking more than one of the small boxes at the front of each option. At times, the number of options which needs to be chosen is stated, at other times students need to choose all of the items which are aligned to a certain characteristic. To register a change, a choice must be unclicked.
Hot spot or hot text	An item response type involves clicking on one or more options within a sentence, passage or image in order to provide an answer. The area(s) clicked are generally highlighted to show that the answer has been registered. To register a change, a choice must be unclicked.
Drag and drop	An item response type in which one or more numerals, words, symbols, or graphics are dragged from one area to one or more designated spaces. Drag and drop items can take several forms. The Drag and drop item forms being proposed for the online NAPLAN tests include: • Position order • Interactive order • Interactive match • Interactive gap match • Interactive graphic match • interactive graphic gap match
Inline choice /drop down menu	An item response type which involves clicking on an arrow next to a text field to reveal a range of available options within the drop down menu, and then touching or clicking the chosen option.
Short answer response	An item response type which requires a word, numeral or symbol to be typed into a text box.
Extended response	An item response type which requires a sentence or more to be typed into the text box provided.
Composite	An item response type which involves two or more different item response types in the one item.

Appendix B
Examples of item types

Type 1
Multiple select



Josephine has 75 seedlings.

She plans to plant them in groups of 15.

Select **all** of the number sentences where the empty box shows the number of groups of 15 she could plant.

☐ 75 + 15 = □

☐ 75 + □ = 15

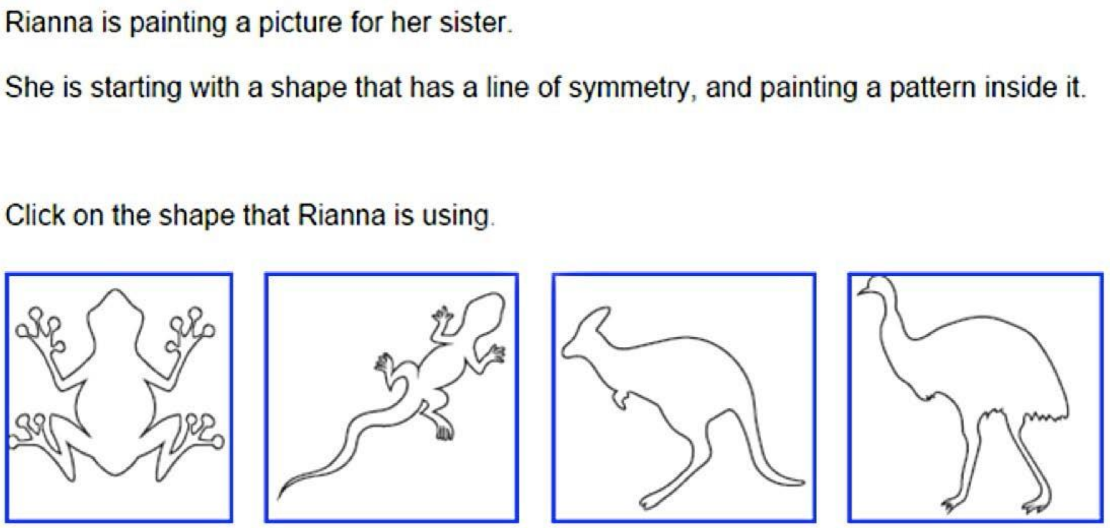
☐ □ + 75 = 15

☐ 75 × □ = 15

☐ 15 × □ = 75

☐ 75 × 15 = □

Type 2
Hot spot

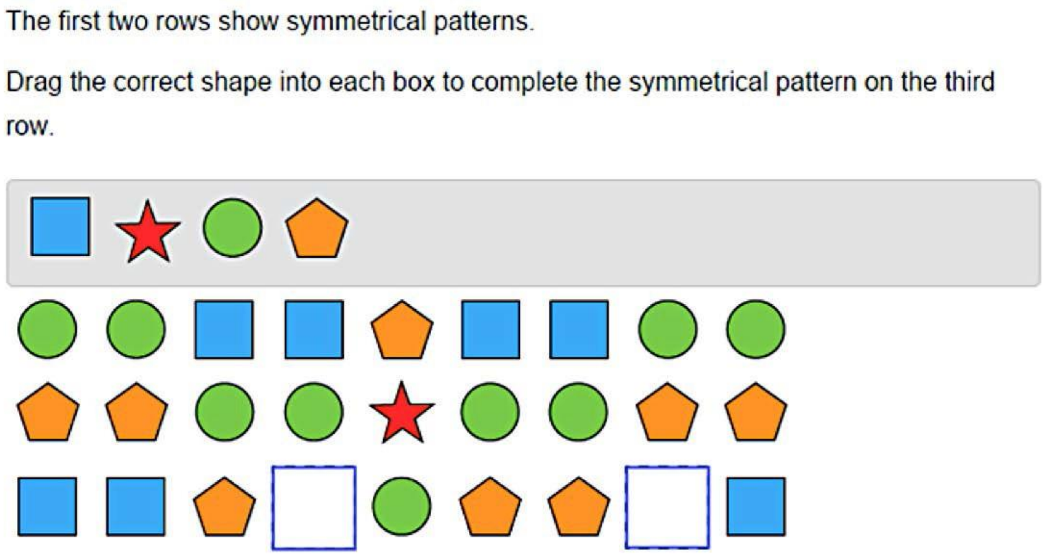


Rianna is painting a picture for her sister.

She is starting with a shape that has a line of symmetry, and painting a pattern inside it.

Click on the shape that Rianna is using.

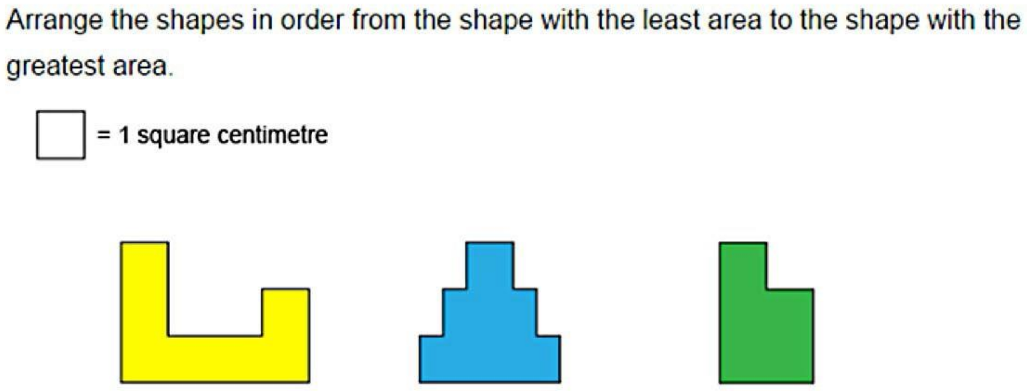
Type 3a
Drag and drop: Position order



The first two rows show symmetrical patterns.

Drag the correct shape into each box to complete the symmetrical pattern on the third row.

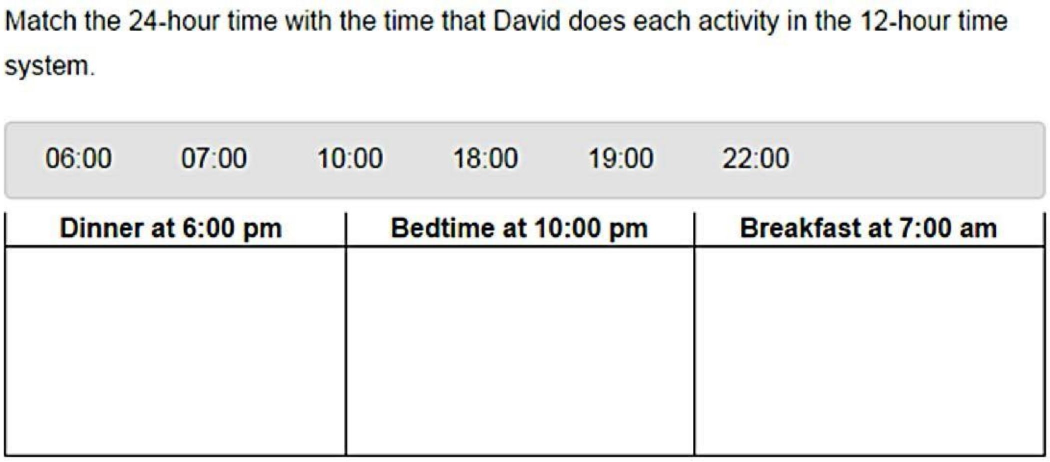
Type 3b
Drag and drop: Interactive order



Arrange the shapes in order from the shape with the least area to the shape with the greatest area.

□ = 1 square centimetre

Type 3c
Drag and drop: Interactive match

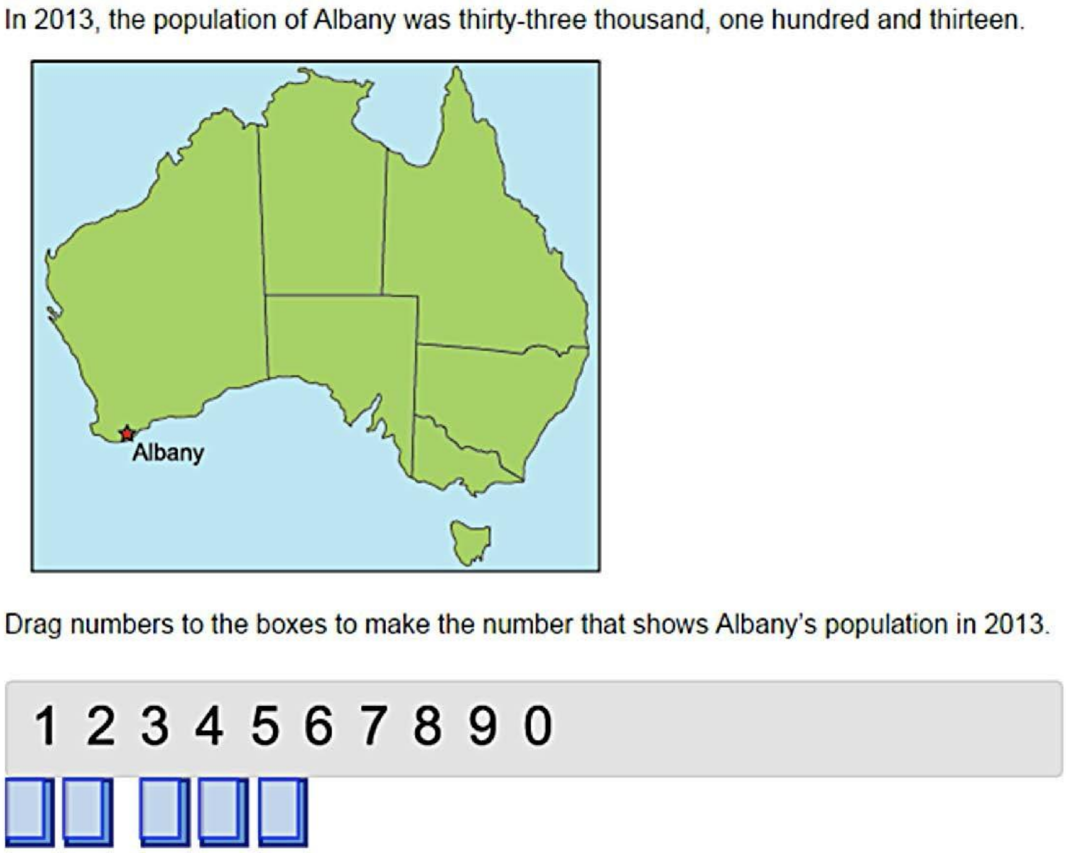


Match the 24-hour time with the time that David does each activity in the 12-hour time system.

06:00 07:00 10:00 18:00 19:00 22:00

Dinner at 6:00 pm	Bedtime at 10:00 pm	Breakfast at 7:00 am

Type 3d
Drag and drop: Interactive gap match



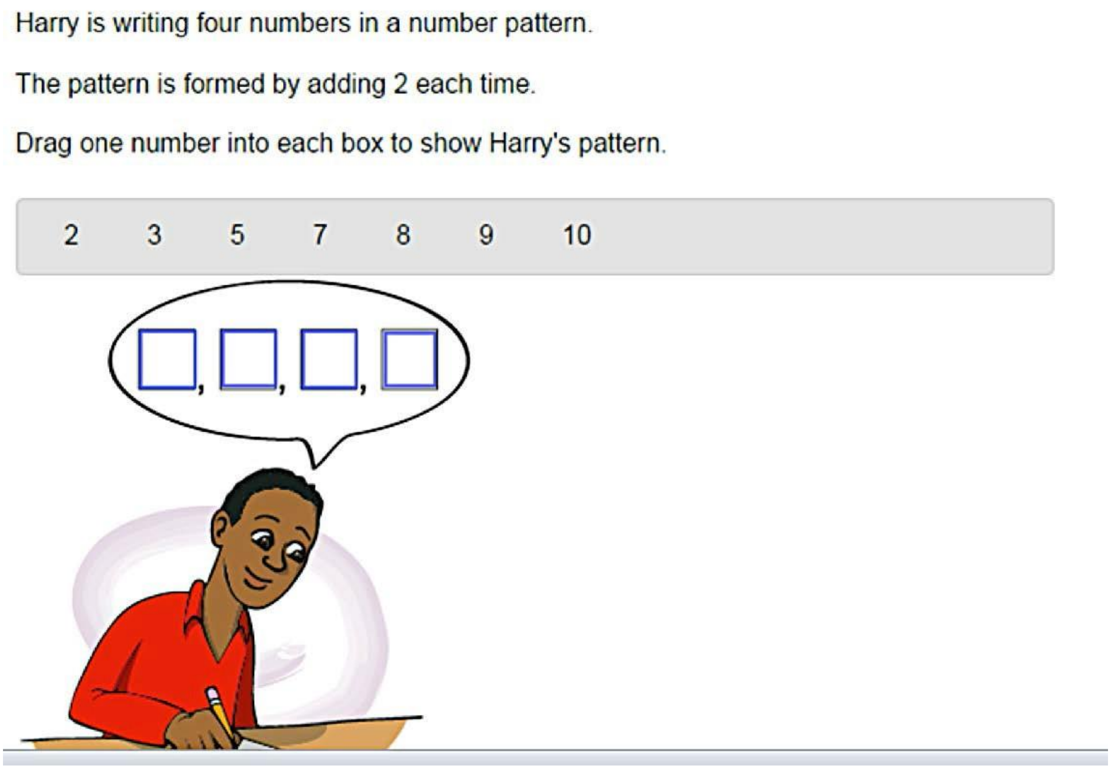
In 2013, the population of Albany was thirty-three thousand, one hundred and thirteen.

Drag numbers to the boxes to make the number that shows Albany's population in 2013.

1 2 3 4 5 6 7 8 9 0

□□ □□□

Type 3e
Drag and drop: Interactive graphic match



Harry is writing four numbers in a number pattern.

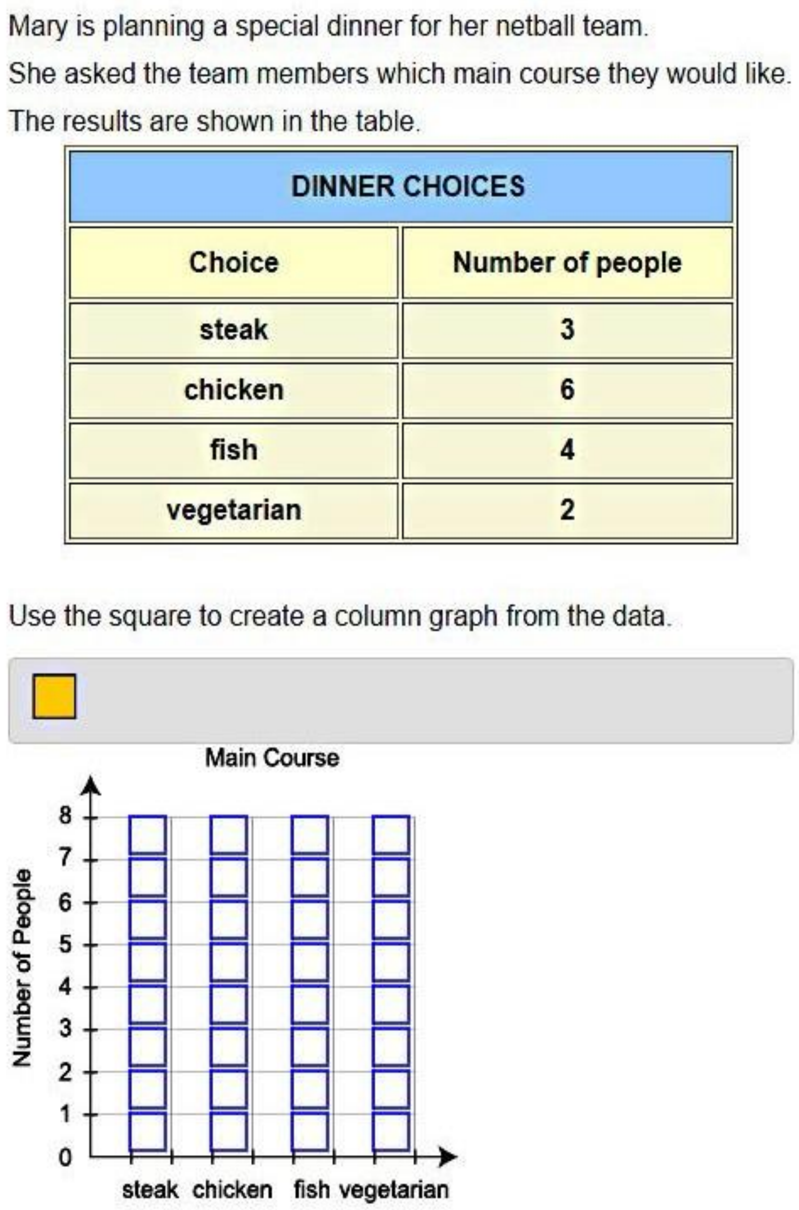
The pattern is formed by adding 2 each time.

Drag one number into each box to show Harry's pattern.

2 3 5 7 8 9 10

□, □, □, □

Type 3f
Drag and drop: Interactive graphic gap match



Mary is planning a special dinner for her netball team.

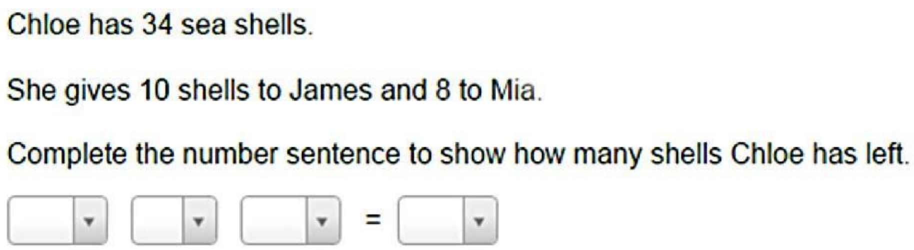
She asked the team members which main course they would like.

The results are shown in the table.

DINNER CHOICES	
Choice	**Number of people**
steak	3
chicken	6
fish	4
vegetarian	2

Use the square to create a column graph from the data.

Type 4
Inline choice/drop down menu

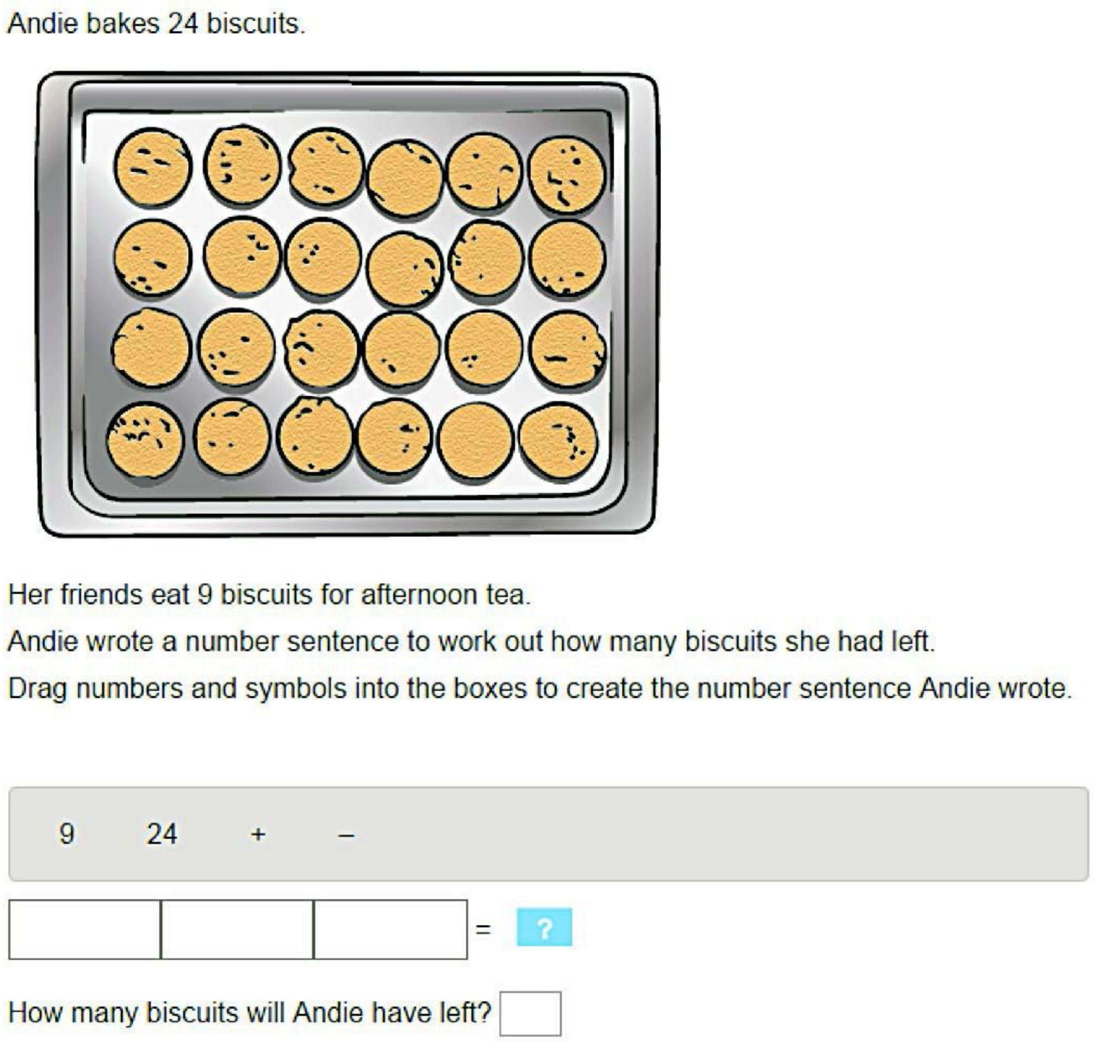


Chloe has 34 sea shells.

She gives 10 shells to James and 8 to Mia.

Complete the number sentence to show how many shells Chloe has left.

Type 5
Composite



Andie bakes 24 biscuits.

Her friends eat 9 biscuits for afternoon tea.

Andie wrote a number sentence to work out how many biscuits she had left.

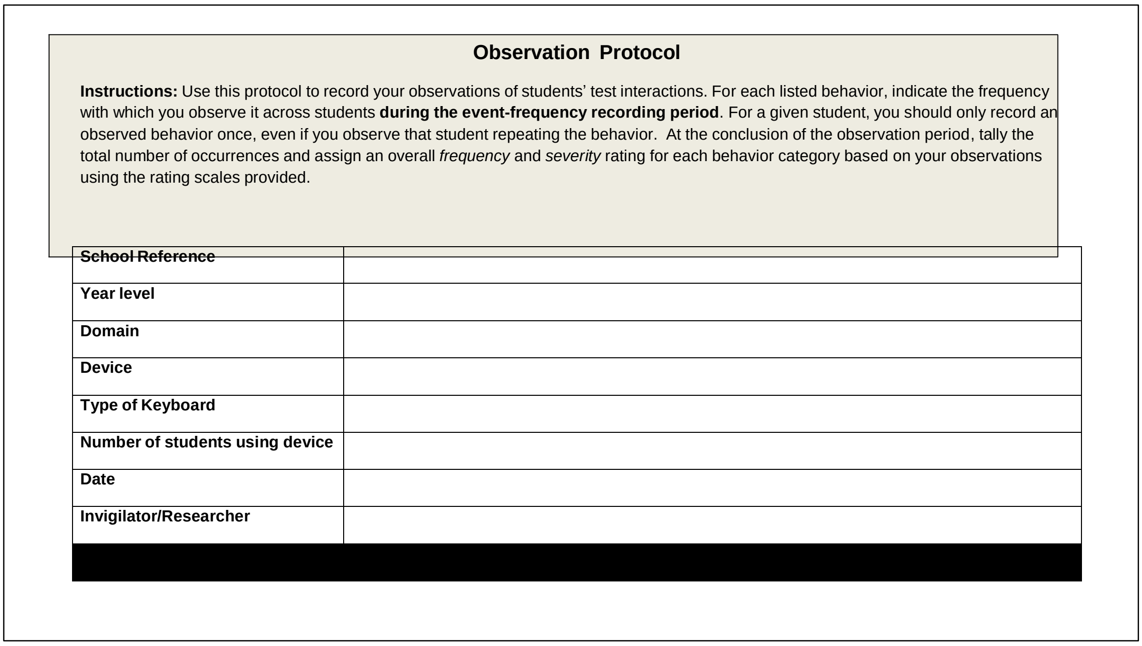
Drag numbers and symbols into the boxes to create the number sentence Andie wrote.

9 24 + –

□ □ □ = ?

How many biscuits will Andie have left? □

Appendix C
Event frequency observation protocol



Observation Protocol

Instructions: Use this protocol to record your observations of students' test interactions. For each listed behavior, indicate the frequency with which you observe it across students **during the event-frequency recording period**. For a given student, you should only record an observed behavior once, even if you observe that student repeating the behavior. At the conclusion of the observation period, tally the total number of occurrences and assign an overall *frequency* and *severity* rating for each behavior category based on your observations using the rating scales provided.

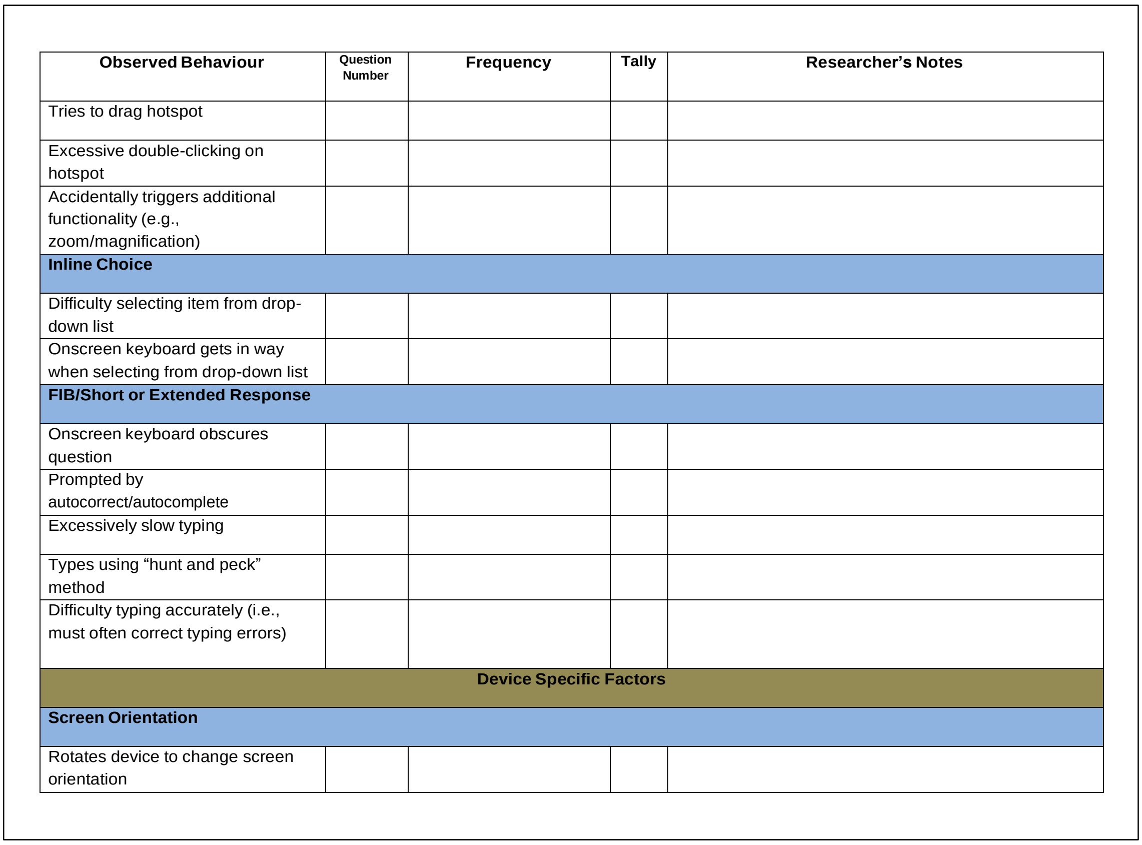
School Reference	
Year level	
Domain	
Device	
Type of Keyboard	
Number of students using device	
Date	
Invigilator/Researcher	

Observed Behaviour	**Question Number** ***if relevant***	**Frequency**		**Researcher's Notes**
		Use hash marks for individual occurrence of observed behaviors, then use that information to determine overall scale rating	**Tally**	

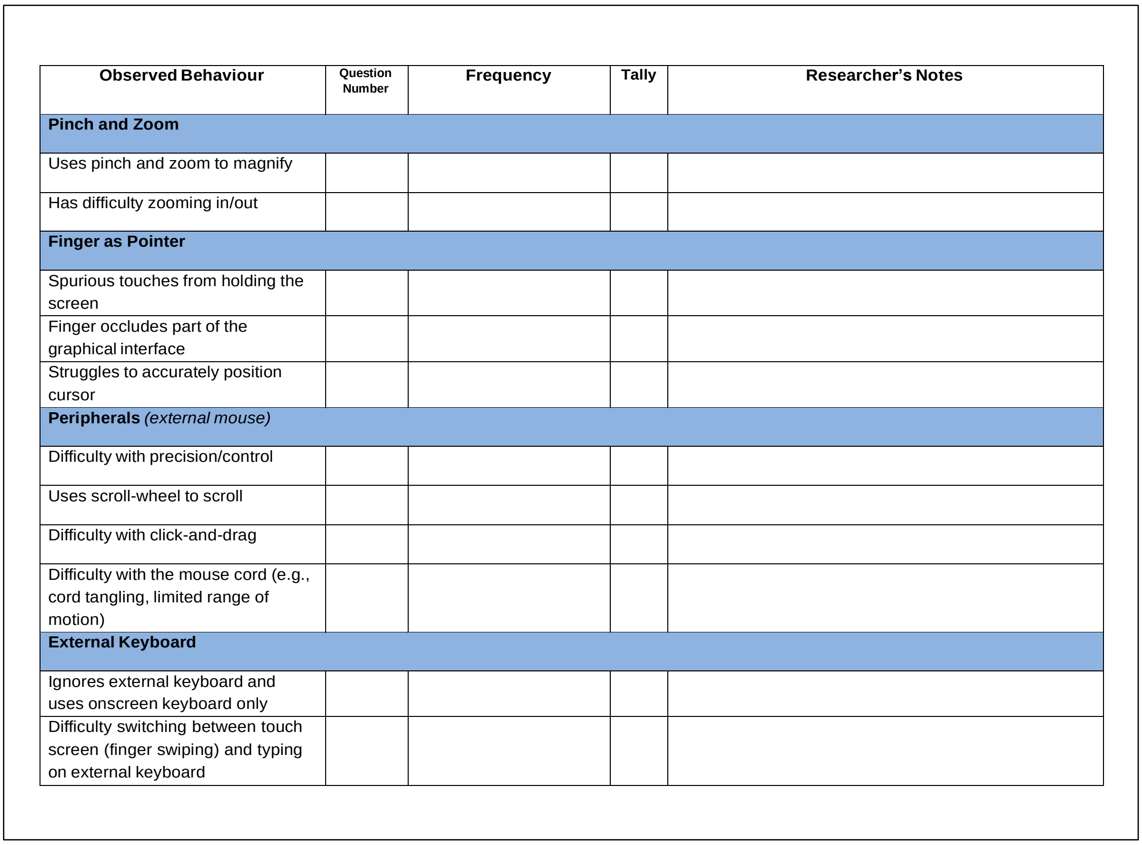
Observed Behaviour	**Question Number**	**Frequency**	**Tally**	**Researcher's Notes**
General				
Navigation				
Confusion with ellipsis				
Confusion related to identifying completed or non-completed items				
Difficulty finding the "Next" or "Previous" button				
Failure to notice that the next item has advanced				
Accidentally exits testing environment during test				
Subject-Specific Factors				
Passage				
Unable to see question and passage at same time				
Difficulty using the arrow button to show/hide the passage				

Observed Behaviour	Question Number	Frequency	Tally	Researcher’s Notes
Difficulty using the “Show/Hide Resource” button to show/hide the passage				
Difficulty scrolling to read the passage (e.g., scrolling in the wrong direction)				
Excessive scrolling from top to bottom to read/respond to the item				
Does not scroll to access text below current screen				
Enlarging or zooming the text in the passage				
Audio				
Difficulty playing or replaying the audio				
Volume control issues				
Replays audio excessively				
Item-Specific Factors				
Viewing Question Stem/Answer Choices				
Difficulty viewing item stem at the same time as response options				
Multiple Choice				
Difficulty with selecting item (cannot engage radio button)				

Observed Behaviour	Question Number	Frequency	Tally	Researcher’s Notes
Multiple Select				
Does not select more than one response				
Drag & Drop				
Finger obscures dragger				
Failure to understand how to change answer (drag out of target)				
Attempts to drag the drop bays				
Difficulty dropping dragger into targeted drop zone (e.g., because inadequate space between drop zones)				
Difficulty understanding that drop area expands (doesn’t understand dragger will fit into drop zone)				
Uncertain where to place draggers that must be ordered/sequenced				
Inability to position dragger causes student to question accuracy of response				
Issues with dragging “across the fold”				
Hotspot				
Difficulty selecting/deselecting items				



Observed Behaviour	Question Number	Frequency	Tally	Researcher's Notes
Tries to drag hotspot				
Excessive double-clicking on hotspot				
Accidentally triggers additional functionality (e.g., zoom/magnification)				
Inline Choice				
Difficulty selecting item from drop-down list				
Onscreen keyboard gets in way when selecting from drop-down list				
FIB/Short or Extended Response				
Onscreen keyboard obscures question				
Prompted by autocorrect/autocomplete				
Excessively slow typing				
Types using "hunt and peck" method				
Difficulty typing accurately (i.e., must often correct typing errors)				
Device Specific Factors				
Screen Orientation				
Rotates device to change screen orientation				



Observed Behaviour	Question Number	Frequency	Tally	Researcher's Notes
Pinch and Zoom				
Uses pinch and zoom to magnify				
Has difficulty zooming in/out				
Finger as Pointer				
Spurious touches from holding the screen				
Finger occludes part of the graphical interface				
Struggles to accurately position cursor				
Peripherals *(external mouse)*				
Difficulty with precision/control				
Uses scroll-wheel to scroll				
Difficulty with click-and-drag				
Difficulty with the mouse cord (e.g., cord tangling, limited range of motion)				
External Keyboard				
Ignores external keyboard and uses onscreen keyboard only				
Difficulty switching between touch screen (finger swiping) and typing on external keyboard				

Observed Behaviour	Question Number	Frequency	Tally	Researcher's Notes
Excessive typing errors				
Difficulty finding numeric keys				
Onscreen Keyboard				
Difficulty opening/closing keyboard				
Keyboard unexpectedly pops-up				
Keyboard obscures item content				
Difficulty navigating to different keyboards (e.g., alpha, numeric, symbols)				
Excessive typing errors				
Ergonomics				
Device Position				
Props tablet up				
Lays tablet flat on table				
Difficulty coordinating external keyboard with tablet position				
Physical Strain				
Appears to experience screen glare or eye strain during testing (e.g., removes glasses, excessive squinting at the screen)				

Observed Behaviour	Question Number	Frequency	Tally	Researcher's Notes
Appears to experience neck or back strain during testing (e.g., excessive stretching of neck or back during testing)				
Tablet Stand/Case				
Issues stabilizing tablet with stand				
Configuration for Testing				
Inadequate workspace to spread out physical materials				

Additional Comments:	
*Please record any noteworthy events, contextual issues, and isolated behaviours which you believe are relevant to the research.	
[redacted]	[redacted]

Frequency & Severity Rating Scales

Instructions: Use the totals from the frequency tallies in the observation protocol to assign a frequency and severity rating for each of the categories.

Frequency Rating	Severity Rating
1. No issues with navigation observed 2. Navigation issues observed for only a limited number of students (e.g., around one-quarter 3. Navigation issues observed for a moderate number of students (e.g., around one-half) 4. Navigation issues observed for most or all students	1. **Minor**: Causes some hesitation or slight irritation. 2. **Moderate**: Causes occasional task failure for some users; causes delays and moderate irritation. 3. **Critical**: Leads to task failure. Causes user extreme irritation.

Perceived ease of use with:	**Frequency Rating** 1. No issues with navigation observed 2. Navigation issues observed for only a limited number of students (e.g., around one-quarter 3. Navigation issues observed for a moderate number of students (e.g., around one-half) 4. Navigation issues observed for most or all students	**Severity Rating** 1. **Minor**: Causes some hesitation or slight irritation. 2. **Moderate**: Causes occasional task failure for some users; causes delays and moderate irritation. 3. **Critical**: Leads to task failure. Causes user extreme irritation.
General		
Navigation *… navigation features*		
Subject-Specific Factors		
Passage *… accessing/viewing passages and answering questions related to passages*		
Audio *… questions that use audio*		
Item-Specific Factors		
Viewing Question Stem/Answer Choices *… viewing question and answer choices*		
Multiple Choice *… answering multiple choice questions*		
Multiple Select *… answering multiple select questions*		
Drag & Drop *… answering drag and drop questions*		
Hotspot *… answering hotspot questions*		

Perceived ease of use with:	**Frequency Rating** 1. No issues with navigation observed 2. Navigation issues observed for only a limited number of students (e.g., around one-quarter 3. Navigation issues observed for a moderate number of students (e.g., around one-half) 4. Navigation issues observed for most or all students	**Severity Rating** 1. **Minor**: Causes some hesitation or slight irritation. 2. **Moderate**: Causes occasional task failure for some users; causes delays and moderate irritation. 3. **Critical**: Leads to task failure. Causes user extreme irritation.
Inline Choice *… answering inline choice questions*		
FIB/Short or Extended Response *… answering Fill-in-the-Blank or Short/Extended Response questions*		
Device Specific Factors		
Screen Orientation *… screen orientation*		
Pinch and Zoom *… pinch and zoom features*		
Finger as Pointer *… using finger as a pointer*		
Peripherals *(mouse)* *… peripherals*		
External Keyboard *… an external keyboard*		
Onscreen Keyboard *… the onscreen keyboard*		
Ergonomics		
Device Position *… the position/orientation of device*		

Perceived ease of use with:	**Frequency Rating** 1. No issues with navigation observed 2. Navigation issues observed for only a limited number of students (e.g., around one-quarter 3. Navigation issues observed for a moderate number of students (e.g., around one-half) 4. Navigation issues observed for most or all students	**Severity Rating** 1. **Minor**: Causes some hesitation or slight irritation. 2. **Moderate**: Causes occasional task failure for some users; causes delays and moderate irritation. 3. **Critical**: Leads to task failure. Causes user extreme irritation.
Tablet Stand/Case *Perceived issues with stands/cases*		
Physical Strain *Perceived issues with physical strain*		
Configuration for Testing *Perceived issues with configuration*		