Leaving Certificate Computer Science

Factors to consider when developing a Computer-Based Examination

Authored by: Paula Lehane

CARPE
Centre for Assessment Research, Policy and Practice in Education

DCU

NCCA
An Chomhairle Náisiúnta Curaclaim agus Measúnachta
National Council for Curriculum and Assessment

Leaving Certificate Computer Science

Factors to consider when developing a Computer-Based Examination

Authored by: Paula Lehane

January 2019

The views expressed in this report are those of the author and do not necessarily reflect the views or policy of the National Council for Curriculum and Assessment.

CARPE Centre for Assessment Research, Policy and Practice in Education

DCU

NCCA An Chomhairle Náisiúnta Curaclaim agus Measúnachta National Council for Curriculum and Assessment

Table of Contents

List of Tables

List of Figures

List of Acronyms

CBEs	Computer-Based Examinations
CS	Computer Science
DES	Department of Education and Skills
eMCAT	electronic Mathematics Common Assessment Task
HCI	Human-Computer Interaction
IDE	Integrated Development Environment
MCQs	Multiple Choice Questions
NCCA	National Council for Curriculum and Assessment
NCEA	National Certificate of Educational Achievement
NZQA	New Zealand Qualification Authority
PARCC	Partnership for Assessment of Readiness for College and Careers
PBEs	Paper-Based Examinations
PISA	Programme for International Student Assessment
SEC	State Examination Commission
UI	User Interface

1.0 Executive Summary Infographic

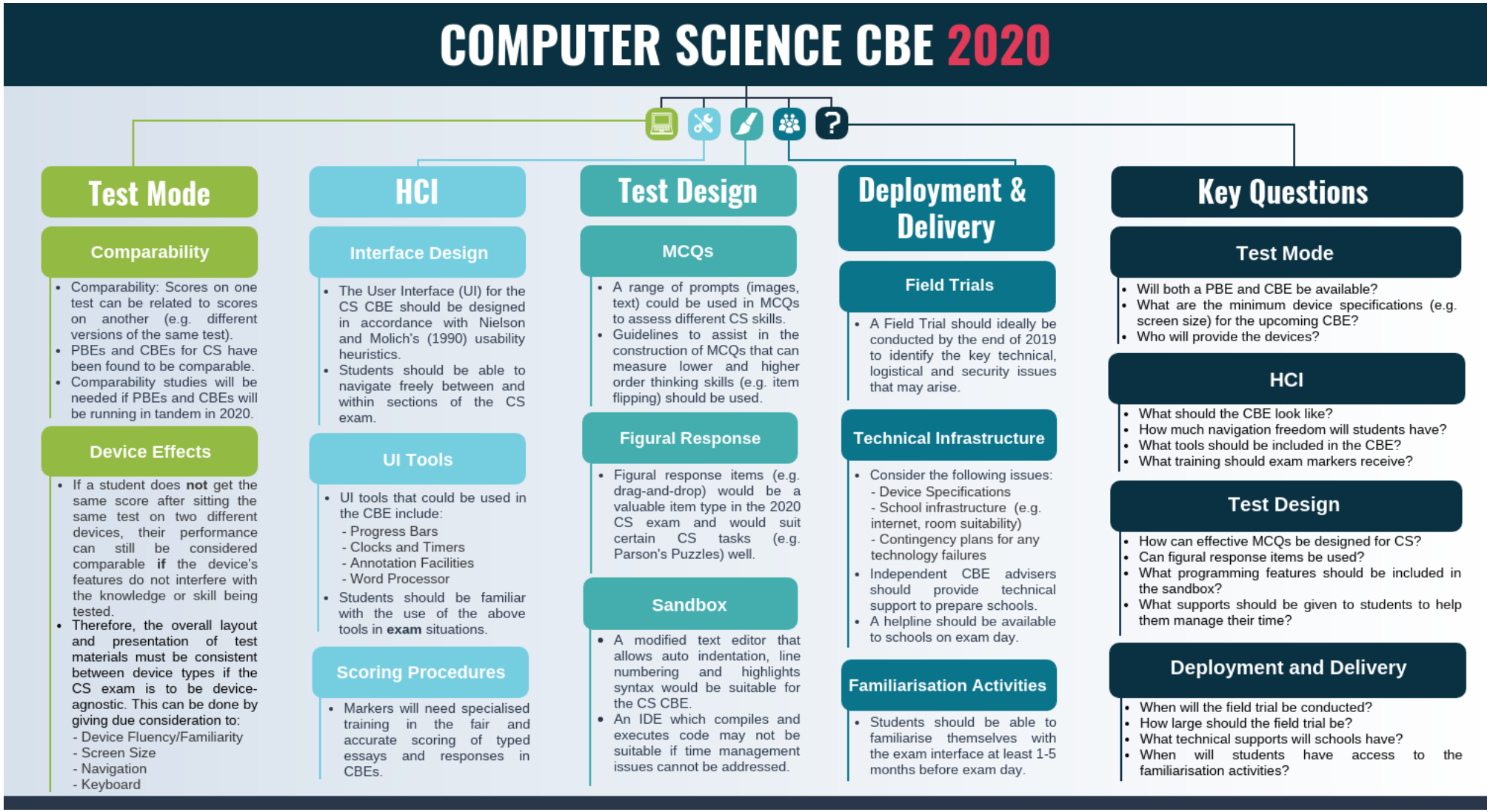


Figure 1 Executive Summary Infographic

2.0 Introduction

The *Digital Strategy for Schools* (Department of Education and Skills [DES], 2015, p. 5) provided a rationale and plan for the embedding of digital technology in all schools in order 'to enhance the overall quality of Irish education' between 2015 and 2020. A more formal approach to the study of technology and computing in second-level schools was advocated in this document which has since resulted in the development of an optional Computer Science (CS) curriculum for Leaving Certificate students (National Council for Curriculum and Assessment [NCCA], 2018). In September 2018, forty schools were selected to trial the implementation of this subject which will culminate in an 'end-of-course computer-based examination' in 2020 (NCCA, 2018, p. 24). This examination will represent 70% of a student's overall CS grade. As achievement in this exam will contribute to a student's overall success in the Leaving Certificate, which can then mediate the future courses of study available to them in further and higher education, it is essential that this exam is appropriately designed and deployed.

The use of a computer-based exam (CBE) for the assessment of CS students is a significant departure in tradition for the Leaving Certificate Established programme. All other subjects in the Leaving Certificate involving an end-of-course examination employ paper-based tests. The planned CBE for CS will represent the first of its kind in the Irish education system when it is introduced in 2020. This challenge of developing and delivering a high-stakes CBE is also magnified by the inherent difficulties associated with the evaluation of students' knowledge and learning in computing courses (Kallia, 2018). Therefore, to ensure that the pending CS exam delivers a CBE in a responsible manner that preserves the fairness, validity, utility and credibility of the Leaving Certificate examination system, several factors pertaining to the design and development of this CBE will need to be considered. In particular, findings from peer-reviewed academic journals within the realm of assessment, human-computer interaction and computer science education and the grey literature of unpublished manuscripts and technical reports from testing organisations will be important to review, as will experiences from countries whose education systems currently use CBEs in high-stakes settings. The current report will discuss the key issues arising from this literature under four broad headings: **Test Mode Comparability**, **Human-Computer Interaction**, **Test Design** and **Test Deployment and Delivery**. Best practice guidelines will be presented at the end of each

section to support the relevant stakeholders in their development of the CS CBE. For readers who wish to reflect on the key messages of the presented literature and how it should inform the future implementation of the planned CBE for CS, Section 8 contains an overview of the key questions that should be considered as a result of the information contained in this report. To begin however, a brief overview of CBEs, and in particular their deployment in CS, will be outlined.

3.0 Computer-Based Examinations (CBEs): An Overview

High-quality assessment is one of the most critical dimensions of an effective educational system (NCCA, 2007). While assessment can take many forms, end-of-course high-stakes exams are often the dominant form of assessment in many second-level systems (Keane & McInerney, 2017). Given the widespread proliferation of digital technology in everyday life, it is unsurprising that many countries are beginning to use high-stakes CBEs to assess their second-level students such as the Partnership for Assessment of Readiness for College and Careers [PARCC] in the United States and the National Certificate of Educational Achievement [NCEA], in New Zealand. For example, the relevant authorities in New Zealand aim to have all NCEA exams for second-level students available online by 2020, thus ending a seven-year transition project from paper-based assessments to computer-based assessments. The PARCC tests for elementary, middle and high school students can be administered by computer or by pen-and-paper according to the relevant state's guidelines. Therefore, Irish efforts to develop a CBE for CS are consistent with current approaches in educational assessment. However, it is important to understand why CBEs are beginning to replace traditional test formats and what outstanding concerns about them as an assessment approach remain.

The primary purpose of any question in a CBE or paper-based examination (PBE) is to 'collect evidence of the test-taker's development of the targeted knowledge, skill or ability' (Russell, 2016, p. 21). Yet, Messick (1988, p. 34) notes that 'tests are imperfect measures… because they either leave out something that should be included . . . or else include something that should be left out, or both'. PBEs are often limited to short answer, multiple-choice or essay-type questions. Therefore, it can be argued that they often 'leave

things out'. Certainly, research consulted by Bryant (2017) claims that PBEs can test only a limited range of knowledge, skills and abilities (also called *constructs* in assessment literature). In contrast, Parshall and Harmes (2008) assert that test items in CBEs can present more authentic contexts for test-takers to demonstrate their knowledge, skills and abilities, thus expanding the test's ability to represent the targeted construct. This is achieved by the greater variety of stimulus pieces (e.g. images, videos, animations) and response options (e.g. drag-and-drop) available in CBEs.

Although CBEs are thought to improve the assessment capabilities of previously established tests and can offer increased opportunities for the assessment of 'hard-to-measure' skills, their overall worth to educational assessment is still under investigation. While there is a 'broad faith' amongst educationalists that CBEs can improve assessments, the exact nature of this value, if it even exists is difficult to verify and describe (Bryant, 2017, p. 1). In particular, there are some concerns that CBEs may introduce construct-irrelevant variance to tests. This is where the actual design or underlying requirements of the CBE can interfere with the knowledge, skill or ability being measured. While every test can potentially have some construct-irrelevant variance, the type of construct-irrelevant variance that CBEs can introduce is slightly different. For example, it can be argued that many PBEs in the Leaving Certificate have some level of construct-irrelevant variance as they over-emphasize the importance of written communication and language, which can advantage or disadvantage different groups of students e.g. those with English as an Additional Language etc., In the case of CBEs, other sources of construct-irrelevant variance exist. If a test-taker cannot fully demonstrate their knowledge in a discursive response about a key figure in computing as a result of their poor typing skills, construct-irrelevant variance has been introduced to the testing scenario. Something unrelated (typing proficiency) has interfered with the measurement of the targeted construct (student knowledge). CBE design (e.g. the ability to review questions etc.,) may also be a source of construct irrelevant variance (Cantillon, Irish & Sales, 2004). Furthermore, some research has found that the type of device used in a CBE can influence the assessment of the intended construct (e.g. Davis, Janiszewska, Schwartz & Holland, 2016).

Despite this, CBEs continue to be promoted in educational systems worldwide as they have a number of important practical advantages compared to traditional PBEs (Csapó, Ainley, Bennett, Latour & Law, 2012). Csapó et al. (2012) state that it takes less

time for examiners to prepare, distribute and score CBEs, thus making them more efficient than PBEs. More detailed analytics on student performance can also be calculated. Furthemore, CBEs allow for more accessibility features (e.g. text-to-speech, increasing font size), thus ensuring that a wider variety of test candidates can interact with the test's content (DePascale, Dadey & Lyons, 2016). Researchers have also begun to develop simulation-based assessments that can be incorporated into CBEs. Such assessments have found to foster higher levels of motivation and engagement in test-takers (Shute, Wang, Greiff, Zhao & Moore, 2016).

However, there are some practical issues that can limit the successful implementation of a CBE. For example, technological infrastructure must be extensively tested and secured in schools before a CBE can be put in place (Haigh, 2010). Recently, several students who sat NCEA digital exams last year in New Zealand erroneously failed some of their tests as a result of a technical 'glitch' (Education Central, 2018). No marks were allocated to some students in the 2017 digital pilot examinations in Classical Studies, English and Media Studies. Similarly, concerns surrounding test-security and the reliability of back-up procedures in case of technological failure for CBEs are also prevalent in literature (Cantillon et al., 2004) and is often a major concern for students (Walker & Handley, 2016). Time for staff and students to get acquainted with new technology is also required which can further increase the already inflated costs associated with the initial development of a CBE (Boevé, Meijer, Albers, Beetsma & Bosker, 2015). Reluctance amongst test-takers to engage with CBEs as a replacement for PBEs has been noted in some research studies as well (Jimoh, Kehinde Shittu & Kawu, 2012).

The strengths, weaknesses, opportunities and threats ('SWOT') CBEs pose to the assessment process, as they stand in literature at present, and in relation to CS, have been summarised in Figure 2 (overleaf).

Figure 2 'SWOT' Analysis of CBEs

3.1 CBEs for CS

Despite the marked preference CS students often demonstrate towards CBEs, anecdotal evidence and published literature indicates that PBEs are still used in introductory computer programming courses in third-level institutions (Barros, 2018; Öqvist & Nouri, 2018). While this persistence with PBEs is likely due to a number of factors including tradition, test security and cost, CBEs for third-level CS courses are beginning to be developed in response to criticisms that summative handwritten exams are not the best way to assess a student's programming ability (Piech & Gregg, 2018; Öqvist & Nouri, 2018). Shuhidan, Hamilton and D'Souza's (2010) study of instructor perspectives for third-level CS courses found that only half of the total respondents (n=66) felt that a summative handwritten programming exam was a valid measure of a student's programming ability. Bennedsen and Casperson (2007, p. 189) explain this perspective by arguing that a pen-and-paper exam is an 'an artificial situation and therefore insufficient and inappropriate to test the student's ability to develop programs'. Certainly, there does appear to be a perception in recently released literature (e.g. Barros, 2018; Öqvist & Nouri, 2018) that handwritten exams can sometimes lack the ecological validity needed to make accurate judgements about a student's programming skills.

Öqvist and Nouri (2018) also highlight a number of other reasons that justify the transition to CBEs for summative CS exams. Student retention is a significant concern amongst instructors in third-level CS courses, which is often contributed to the difficulties associated with learning to program (Sarpong et al., 2013). Winslow (1996) claims that it generally requires up to 10 years of experience for a novice to become an adept programmer, and assertion supported by Gross and Powers (2005). This may have significant implications for the Senior Cycle CS subject. While this declaration strengthens the argument that CS should be introduced to students at a younger age, it also means that any Leaving Certificate CS exam should be carefully devised. Realistic expectations of what can be achieved in a two-year course must be formed and any assessments should take into consideration that Leaving Certificate students are only beginning to develop their programming skills. Öqvist and Nouri (2018) state that using PBEs for the assessment of developing programming skills is inadequate in introductory programming courses as it fails to take into consideration how *novice and beginner* students learn and apply their recently acquired programming skills. For example,

memory retrieval is an important predictor of success in programming assessments as students must be able to recall information about syntax, algorithms and design patterns (Thompson et al., 2008). When students learn this information they do so with access to a computer with an Integrated Development Environment (IDE) as well as a myriad of other resources. While it is not possible to allow students access to all possible resources in an exam, assessing programming ability with a pen-and-paper exam ensures that there is a significant difference between the learning and assessment environments. This could have a negative impact on student performance as there are no contextual cues to support a novice programmer's recall of key information (Öqvist and Nouri, 2018). There is also a risk that the learning environment would evolve to mirror the pen-and-paper mode of assessment, which would undermine the CS specification. Certainly, Gipps (1994, p. 29) asserts that to ensure the accurate assessment of student learning, assessment procedures should be closely related to 'the models that students construct for themselves'. This can be seen in other Leaving Certificate subjects (e.g. Physics), where students are allowed to consult formulae in log tables, a common classroom practice, during exams. If student learning occurs on computers, then so too should student assessment. Therefore, if student learning occurs in a technology-supported environment, then there is a clear argument for the use of CBE in assessment.

CBEs are also thought to be more supportive of novice programming strategies (Öqvist & Nouri, 2018). Given the amount of time and practice required to attain a high level of proficiency with computer programming, assessment approaches should take into consideration the expected skills of the target assessment group. Programming is an iterative process involving these steps according to Lister et al. (2004):

1. Abstracting the problem
2. Generating sub-problems
3. Composing sub-solutions
4. Recompose
5. Evaluate and Iterate

Applying these steps will require students to rearrange and modify their thoughts and their written code throughout the problem solving process. This approach to problem-

solving process is not supported by a PBE, as students cannot easily edit their code. Furthermore, Winslow (1996) found that novice programmers tend to examine and construct code using a line-by-line approach. Unlike expert or intermediate programmers, they do not plan their solutions in advance or in meaningful chunks. This is a requirement when completing handwritten programming tasks. To be able to properly assess novice students' programming ability, these strategies need to be taken into account when choosing an appropriate exam mode for novice programmers.

As beginner programmers will be assessed in the Leaving Certificate in 2020, the use of a CBE for a second-level CS exam is both reasonable and forward thinking. However, discussions over the level of functionality that should be included in a CBE for CS are still ongoing. Rajala et al. (2016) note that CBEs for CS can provide students access to appropriate programming environments that will allow them to write and edit a code using a keyboard. Editors that take care of code indentation and syntactic colour coding can be included as can IDEs that can compile and execute code. While features such as these are standard when learning to code, their inclusion in an examination setting may interfere with the assessment process (Dillon, Anderson & Brown, 2012). Therefore, while CBEs can offer many educational and practical advantages over PBEs, their design and implementation can be challenging. Test developers for the upcoming CS exam should strive to gain a thorough understanding of CBEs to ensure that a coherent and integrated strategy for their reliable development and secure deployment for a CS context is guaranteed.

4.0 Test Mode Comparability

When test variations occur due to the use of different administration formats or devices for the same test, guidelines need to be in place to ensure that the test is still **comparable** (Winter, 2010). Therefore, evidence of comparability is required between pencil-and-paper assessments (e.g. PBEs) and technology-based assessments (e.g. CBEs), as well as between the different devices a technology-based assessment is administered on (e.g. tablets and laptops). Depending on the administration procedures that will be in place for the CS exam in 2020, issues in relation to test mode comparability will need to be evaluated.

4.1 PBEs and CBEs

During the initial development of CBEs, discourse surrounding test variations and the importance of comparability between PBEs and CBEs began to dominate research literature (Kingston 2009; Winter, 2010). This subsequently led to several studies by a variety of educational researchers comparing paper-based and technology-based variations of the **same** test. Such studies aimed to determine if mode effects can occur as a result of administration format. Mode effects refer to whether 'questions designed to be delivered on paper are systematically easier or harder when delivered on computer' (Jerrim, 2018, p. 16). Paek's (2005) summary of comparability studies found that out of 97 cases, the results of CBEs and PBEs were comparable in 74 cases; in 8 cases, the computer-based test appeared to be more difficult; and in 15 cases, the paper-and-pencil test appeared to be more difficult. Similarly, Kingston (2009) synthesised the results of 81 comparability studies performed between 1997 and 2007. Based on test scores, CBEs appeared to provide a small advantage for English Language Arts and Social Studies tests (effect sizes of.11 and .15, respectively), and PBEs provided a small advantage for Mathematics tests (effect size of –.06). Another meta-analysis by Wang, Jiao, Young, Brooks and Olsen (2007) also aimed to determine the impact (if any) of the administration mode on mathematics performance in students between K-12 grades in the United States. They found that effect sizes for administration mode on test scores was 'negligible'.

While these studies indicate that, in general, there is no ***practical*** difference between CBEs and PBEs when measuring student achievement and performance, some mode effects between CBEs and PBEs ***can*** occur. In the US, students who took the 2014-15 PARCC exams via computer tended to score lower than those who took the exams with paper and pencil (Herold, 2016). In one state (Illinois), 32% of the 107,067 high-schoolers who took the test online were deemed proficient. This contrasts with the 17,726 high school students who took the paper version of the exam where the percentage of proficient students was found to be much higher (50%) (Illinois State Board of Education, 2015). While this pattern was noted in other states and districts, it was not consistently found in all cases (reported in Herold, 2016). Reasons for these score discrepancies were unclear and caused significant debates in educational circles in the US, with many querying the validity and reliability of the PARCC assessments (Herold,

2016). According to Kingston's (2009) and Wang et al.'s (2007) research, differences in test-taker performance between CBEs and PBEs, if they occur, can depend on a number of factors including subject area, test-taker demographics (e.g. age, computer skills) and differences in question presentation and response format (refer to Figure 3). As a result, mode effects are somewhat difficult to predict.

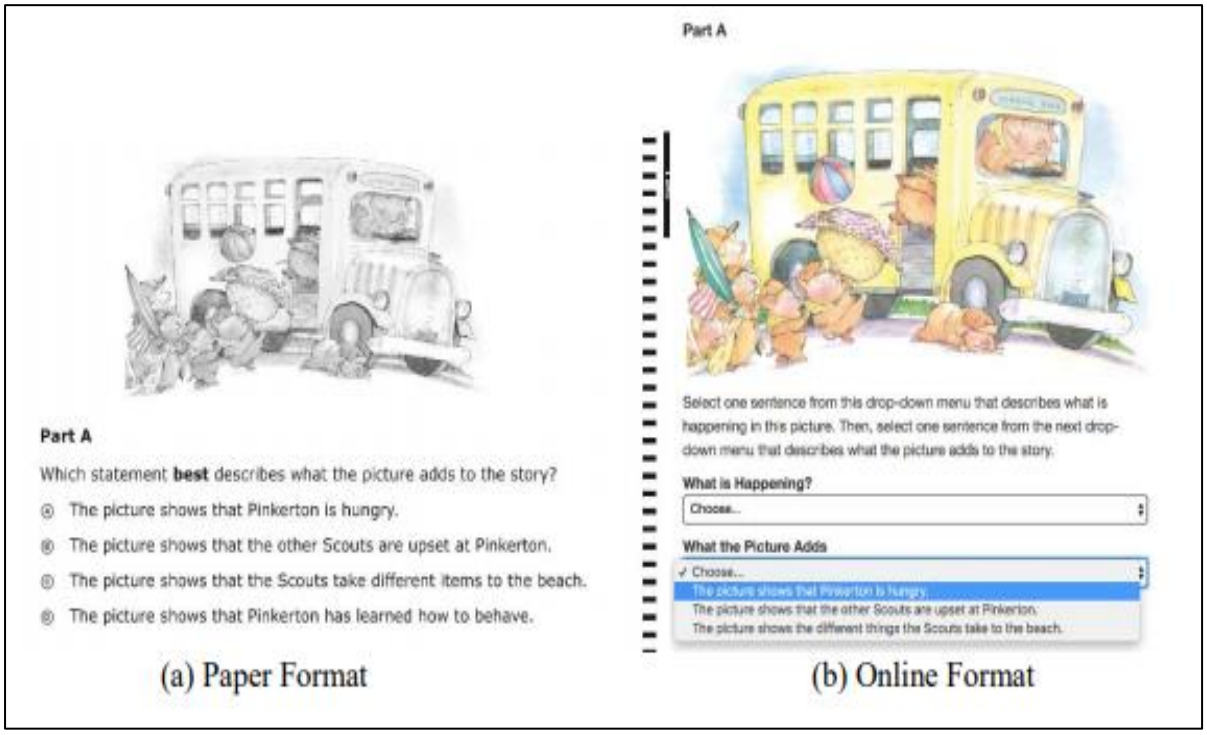


Figure 3 PARCC sample test (obtained from Backes & Cowan, 2018)

When different versions of a state assessment are administered (e.g. CBEs and PBEs) in the US, the state must be prepared to provide results of a comparability study that provides 'evidence of comparability generally consistent with expectations of current professional standards' (United States Department of Education, 2015, p. 43). This is derived from international best practice guidelines which assert that evidence of comparability should be provided whenever there are variations in the content of an assessment or its administration. This explains the origin of the previously mentioned comparability studies where content-equivalent PBEs and CBEs were available to test-takers. Unfortunately, very little high quality research that directly investigates the comparability of PBEs and CBES for CS exist. Öqvist and Nouri's (2018) study aimed to identify the presence of any mode effects in the performance of novice programmers taking an identical CS exam on a computer with minimal programming support (e.g. a

text-editor rather than an Integrated Development Environment was used) or using a PBE. While some differences were noted between the two groups, the overall results demonstrated no statistically significant differences between the groups in terms of performance. Yet, it is important to acknowledge that the sample size for this study was quite small (n=20) which calls into question the generalisability of these results. Another study by Gursoy (2016) involving 44 students found that there was also no statistically significant exam score difference between students who took a CBE and PBE exam for an introductory CS course in a third- level institution. This research indicates that computer-based CS exams, **if** carefully designed, can be comparable to a paper-based equivalent. Yet, it must be noted that, due to limited research, it is difficult to say this with a high degree of certainty. Mode effects may still occur between paper-based and computer-based CS exams. Still, what research is present appears to indicate well designed CS exams are comparable between different modes of assessment. This may be of some interest to the State Examination Commission (SEC) and the NCCA as it suggests that a paper version of the CS exam ***can*** be used in tandem with a computer based version ***if*** necessary.

At present, it appears that there are no plans to develop both a PBE and a CBE for the upcoming CS exam (Halpin, 2018). If a CBE will be the only test form deployed in 2020, the relevant testing authorities will not have to prove comparability between administration modes. However, if, for practical or logistical reasons, a PBE does run alongside a CBE (e.g. where students can choose to complete a PBE or CBE in June 2020) or if a PBE will be used if a technology failure occurs, then data from some form of comparability study should be available to demonstrate that the two test forms are comparable. Other comparability concerns should also be attended to in advance of June 2020. Research (e.g. Dadey et al., 2018) indicates that mode effects can also be found *within* a particular administration format. In the case of the Leaving Certificate CS exam, if different devices and device types are used to administer the planned CBE then issues relating to device comparability must be deliberated upon.

4.2 Device Comparability

The term 'device' refers to a 'range of technology hardware that can be used to access digital content and can include a wide array of input mechanisms, output mechanisms, shapes and sizes' (Davis, Janiszewska et al., 2016, p. 5). Examples of digital devices that are commonly used for education, training and assessment purposes include smartphones, tablets, eReaders, laptops and desktop computers. Each of these types of devices have what Way, Davis, Keng, and Strain-Seymour (2015) call a *form factor*. Form factor describes the size, style, shape, layout and position of the major functional components of a device which determines how the user engages with the device to access and manipulate digital content (Way et al., 2015). Table 1 (overleaf) from Lehane (2018) identifies some of the key features that contribute to a device's form factor and summarises the variations that occur between devices as well as the possible implications of these for users.

Table 1 Overview of key device features and common variations for each (Lehane, 2018)

Feature	Variations	Implications
Screen	- Size (ranging from 5.5"- 23") - Type (LCD, IPS-LCD etc.,) - Resolution (number of pixels per inch resulting in HD, Ultra HD etc.,)	Content distribution is affected by screen size. Screen types and resolutions have ergonomic implications e.g. visibility in light/dark conditions.
Navigation/ Input	- Touchscreen (finger, stylus) - Peripherals (Mouse, trackpad, Trackball) - Voice control	Accuracy and functionality of each approach varies between devices e.g. Touchscreens do not have a 'pointer' that allows for tracking or hovering.
Text Entry	- Speech-to-Text - Keyboards (external, on-screen)	Speech-to-text accuracy is still unreliable. On-screen keyboards often take up a large amount of screen estate.
Output	- Sound - Lights - Vibrations	Feedback or information can be given to users in several ways which is useful for personalisation.

User experience between desktops and laptops can be expected to be relatively comparable because the form factors of the devices are similar. However, there are significant differences in the form factors between computers and touchscreen tablets that may influence student experience in using the devices. Therefore, it can be asserted that different form factors may cause significant variations in test-taker experience which could potentially impact on test-taker performance as well as the comparability of attainment and ability measures across devices. Schroeders and Wilhelm (2010) identify four ways that differing form factors can impact assessment situations: perceptual demands, motor skill requirements, item presentation and device familiarity. Different devices will cause changes in the presentation of and response to test items due to different screen sizes and input methods that, depending on device familiarity and proficiency, could present significant motor and perceptual challenges. This could have a subsequent impact on performance. As a result, if the CBE for CS is to be device-agnostic (where the CBE can be used on any device), test developers should be aware of the occurrence of device effects in previous studies of high-stakes tests amongst second-level students as well as the ways in which such device effects can be minimised.

4.2.1 Device Effects

Using the terminology contained in work by DePascale et al. (2016), the comparability of scores produced by students taking the same assessment on different devices is referred to as device comparability. This term does not refer to the form factors of the actual devices and to what level they have comparable technological features. Instead, it indicates that the 'scores resulting from the assessment administration on different devices are comparable' (Dadey et al., 2018, p. 31). Research has found that when the same test is taken on two different devices, they may produce the same ***overall*** score distributions but scores for ***individual*** test-takers may vary between devices due to individual differences in device use (Dadey et al., 2018; Lotteridge et al., 2010). Scores from different administration devices can potentially lead to a rank ordering of test-takers that varies by device or even by item type and device, as was seen in a large scale study (n>60,000) conducted on eight PARCC tests administered to high-school students in America (Steedle, McBride, Johnson and Keng, 2016). Similarly, a between groups study by Davis, Kong and McBride (2015) investigated the comparability of scores for

high school students (*n*=964) on 9.7" tablets (*n*=485) and a mix of desktop and laptop computers (*n*=479) with no required specifications. No statistically significant performance differences in student test scores were noted across device types for the content areas of reading, mathematics or science. However, in a secondary analysis using the data set from their 2015 study, Davis, Morrison, Kong and McBride (2017) evaluated device comparability across a range of student ability levels and key demographic variables like gender and ethnicity. Comparable performance across device conditions was noted across many student subgroups except in the reading test where performance was enhanced by the use of tablets for male students. Yet, it is important to note the effect size for this interaction was quite small indicating that while a difference was present, the size of this difference was small from a practical perspective.

The different form factors of devices make it impossible to ensure that test-takers complete CBEs 'under the [sic] tightly specified set of conditions' (DePascale et al., 2016, p. 15) needed for standardisation. Instead, as recommended by Dadey et al. (2018), it is more appropriate for test-developers to assert that a test is functionally comparable in relation to overall performance but that individuals could take the test on a device that is most likely to produce the most accurate estimate of their ability i.e. a device of their own choice. This comparability claim presents an increased level of flexibility. Dadey et al. (2018) warn that this is only appropriate if the content basis of the test variations in relation to the score types produced is kept constant across all devices regardless of form factor variations. Therefore, when adapting assessments for use on multiple devices, it is essential that the construct being measured is not altered by device-driven features.

It appears that some device effects do occur when deeper analyses are conducted on data sets that compare test-taker performance across device types. While the device effects are often small or only occur in specific contexts, it is unclear how they can emerge. To minimise the occurrence of these device effects, research into the potential factors or device features that may contribute to the presence of device effects should be consulted.

4.2.1.1 Device Familiarity/ Fluency

The degree to which a test-taker is familiar with a device and how to use it has been a key concern for test-developers for some time (Lorié, 2015; Davis, Janiszewska et

al., 2016). Many researchers such as Lorié (2015) and Schroeders and Wilhelm (2010) consider test-taker unfamiliarity with a device to be a potential threat to device comparability as failed attempts to access content could lead to test-taker frustration as well as construct irrelevant variance. It can be argued that individuals who study or work with a particular platform or device will be more likely to do better on a test administered on that device or similar. However, device fluency and familiarity can vary by age and preference can also be influenced by content and purpose. In a small-scale study (*n*=24) using cognitive laboratories ('think-aloud') as a methodology, students who were familiar with tablets experienced the same usability issues as those with no experience of tablets (Strain-Seymour, Craft, Davis and Elbom, 2013). However, students with tablet experience handled unintended functionality, such as accidental zooms, better than those with no experience. While all those surveyed thought that the tablet assessment was 'cool', the majority of the students reported that they would prefer to write essays and other discursive responses using desktops rather than tablets.

Form factor appears to play a role in test-taker familiarity, fluency and ease with the device of administration in testing scenarios. The possible effects that device features (e.g. screen size, touchscreen, keyboards etc.,) could have on test-taker experience should be determined. Controls to minimise the effects of these features could then be incorporated into the planned CBE for CS.

4.2.1.2 Screen Size

Bridgeman et al. (2003) investigated the effect of screen size, resolution, internet connection speed, operating system settings, and browser settings for the digital version of the SATs (Scholastic Aptitude Test) for college bound students (*n*=357). While the results showed that some test-takers were frustrated when using small screen sizes with lower resolutions (e.g. 15" screen with 640 x 480 resolution versus 17" screen with 1024 x 768 resolution), only the amount of information available on the screen without needing to 'scroll' had a significant impact on performance in reading subtests. Lower scores were observed in verbal subtests when smaller screen sizes and resolutions led to a lower percentage of the reading materials being visible at one time. Similarly, Sanchez and Goolsbee (2010) found that characters and font sizes that increased the amount of scrolling on screens of any size resulted in lower levels of factual recall.

Research has consistently found that the role of screen size in test-taker performance appears to be influential (Dadey et al., 2018). Performance on recall tasks appears to be influenced by the increased level of scrolling associated with smaller screens. This increase in scrolling appears to introduce an excessive amount of cognitive load in test-takers – a major construct irrelevant variance between devices of different screen sizes according to Sanchez and Goolsbee (2010). Test-taker performance and experience also seems to be hampered if any portion of the screen is blocked, if the reading passage and assessment items are not presented side by side or if a large amount of negative space is present as noted by observations of test-taker behaviours in Davis, Janiszewska et al. (2016) and Strain-Seymour et al. (2013). All of this appears to contribute to a level of cognitive load not associated with the content being assessed. It is possible then that screen size can introduce construct irrelevant variance in the presentation of tests across different devices. This suggests that the presentation of written texts (e.g. vignettes) across multiple device types needs to be carefully considered as poor interface or platform design could result in excessive scrolling. This then contributes to the presence of construct-irrelevant variance.

With this in mind, a 2010 study of middle and high school students by King et al. (2011; *n*=1547) varied the screen sizes that students would use to complete an online assessment. The researchers kept the amount of information shown on screen, the screen resolution, and the amount of scrolling required equal across all test conditions. When these variables remained constant, no difference in student performance using computers with different screen sizes (10.1", 11.6", 14", 21") was noted. These results suggest that the amount of information available on a screen at one time has a greater impact on test-taker performance than the actual size of the information placed on the screen (as long as the content is still legible). This has several implications for the design of user interfaces and platforms that can be used across devices that test-developers should consider which are discussed at the end of this section.

4.2.1.3 Navigation

As mentioned previously, input mechanisms vary greatly across devices. Devices can be controlled by a variety of approaches including voice control, touchscreens and peripheries like a mouse, trackpad or stylus. Given the only recent development of

reliable speech controlled software for devices, device comparability research has focused on the comparability of test-taker performance on the use of devices with and without touchscreen capability. Touchscreen devices facilitate very specific device-user interactions not available on other devices such as pinch-to-zoom magnification, gesture control etc., However, Way et al. (2015) assert that the issue of comparability comes to the fore when it is unclear if these features interfere with the construct being measured.

Dadey et al. (2018) note that the most common issue with fingertip input noted by research studies in this area (e.g. Strain-Seymour et al., 2013; Eberhart, 2015) occurs when objects in the screen requiring interaction (such as drag-and-drop) are close in size or smaller than the student's fingertip, or when objects are close together. Touch-screen input accuracy may also suffer from accidental touches that result from holding the device. Therefore, touch inputs are associated with high speed but reduced precision. in assessments where reaction time is a factor, a touchscreen could introduce some construct irrelevant variance. Furthermore, a mouse or trackpad controlled cursor can be moved without triggering a selection state on a device. Eberhart (2015) points out that hovering the cursor over an unknown icon can provide information on the icon's purpose or can be used as a pointer to guide reader attention in texts. The features associated with laptops and desktops provide students with contextual information that may assist them in navigating an assessment. These 'hover' features are not available on touchscreens.

Although it is difficult to say for certain if this feature of touchscreen devices can introduce construct irrelevant variance, a small scale observational study noted that users who were unfamiliar with different testing interfaces found that the absence of a cursor, as is standard for touchscreens, led to less user feedback for navigating or troubleshooting purposes (Strain-Seymour et al., 2013). For these reasons, it is important to consider the role of touch-screens when designing user interfaces and item interactions.

4.2.1.4 Keyboards

Two broad categories of keyboards exist – external and onscreen. External keyboards are most commonly associated with laptops or desktops but external keyboards are also available for tablets. Onscreen keyboards are launched or dismissed

from the screen as necessary. Onscreen keyboards tend to have multiple versions that can be navigated through to find numeric or symbolic characters, unlike external keyboards which require a combination of key presses for such access. Strain-Seymour et al. (2013) acknowledge that test-taker frustration may emerge if errors occur as a result of the test-taker not knowing how to access certain characters.

When Dadey et al. (2018) reviewed the literature on the topic, it appeared that on-screen keyboards work equally as well as external keyboards for short or single-response items. However, one research study exploring the use of external and onscreen keyboards found that student responses tend to be reduced in length when using onscreen keyboards for responding to open-ended or composition items (Davis, Orr, Kong and Lin, 2015). Yet, Davis, Orr et al. (2015) found that this does not seem to impact on their overall performance or grade – the types of keyboards may change the content of their responses but it does not hamper the *construct* being measured. In their analysis of essays completed by 826 5th and 10th grade students in America, training or familiarity with a particular keyboard type did not impact performance but older students did show a *preference* for external keyboards. This indicates that student preference does not always equate with a change in student performance which may have several design implications for user interfaces.

On-screen keyboards may have an unintentional interaction effect with screen size. An onscreen keyboard takes up screen real estate and may cause scrolling as it blocks part of the question and test-taker response which, as noted by Sanchez and Goolsbee (2010), can have a negative impact on test-taker recall and performance. As stated by Dadey et al. (2018), this does have an impact on test-taker performance as having to switch between screens etc., causes some construct irrelevant variance that interferes with test-taker engagement. This suggests that some caution should be used when interpreting the role of onscreen keyboards in student performance

4.3 Guidelines for CS Examination: Test Mode Comparability

- A CBE is an appropriate assessment mode for the CS exam as this will reflect students' learning experiences in accordance with Gipp's (1994) work.
- There is a limited amount of research that indicates that CS exams are comparable when administered using paper-based or computer-based forms.

- A comparability study between PBEs and CBEs is advised for the 2020 CS exam ***if*** both administration formats are available to students. The study should ensure that scores on the CS exam are functionally comparable regardless of test mode and that any mode effects do not have any practical significance to students.
- Clear guidelines regarding what devices can be used for the examination to minimise the occurrence of any device effects should be provided to test-takers (e.g. minimal screen size etc.,). The presentation of test-content should be kept as consistent as possible across different device formats.
- When investigating the comparability between scores achieved by students on different devices, it is important to consider that **the same skills can be tested on different devices and that students can get a score that best reflects their ability on a device of their choosing** (Dadey et al., 2018) if device features (e.g. screen size etc.,) do not interfere with what being tested.
- To ensure that construct equivalence is present across devices, an awareness of the differences in test-taker experience that specific device features create must be present e.g. if excessive scrolling occurs, pictures appear in different places etc.
- Efforts to control these differences in test-taker experiences between devices should also be undertaken by key stakeholders e.g. locking a test-screen so that it can only be accessed in landscape mode to minimise scrolling etc.
- If lengthy reading passages or vignettes are involved in the CS exam, then careful consideration in relation to the design of display interfaces to maintain legibility, minimise scrolling and avoid the use of tools that block content on devices with small screens should be considered.
- If students will be permitted to use tablets to take the CBE for CS, then an external keyboard may need to be recommended for use based on the research presented here. If onscreen keyboards are used, they should be carefully designed so that they are still functional but without taking up an excessive amount of screen space.
- The use of a touchscreen for certain test items (e.g. drag-and-drops etc.,) could cause some construct irrelevant variance. These should be accounted for when designing tests across multiple devices.

5.0 Human-Computer Interaction

Human-Computer Interaction is the study of how people interact with computers to ensure that they can carry out a designated activity (Preece et al., 1994). By adhering to the recommended guidelines associated with human-computer interaction, users should be able to achieve their specified goals in an effective, efficient and satisfactory manner (Tidwell, 2011). Ensuring the usability of the upcoming CBE for CS should be a key aim for all stakeholders. If the usability of this CBE is low, test-takers may not understand how to use the instrument or perform slowly. Therefore, the user will not be able to adequately demonstrate their level of proficiency and their score may not accurately reflect the user's competency in CS. To avoid this, test-developers should be aware of some of the key research findings surrounding human-computer interaction and how they can be applied to the upcoming CBE for CS in 2020.

5.1 Interface Design

According to Tidwell (2011), user interface (UI) refers to the means through which users interact with the hardware and software of electronic devices. The UI determines how commands are given to the device, how information is displayed on the screen and how the user can navigate the system. Typography, media files, forms, menus and many other elements all contribute to the UI of a device or software. All of these components should be designed in a way that facilitates an easy and efficient user experience. The most commonly cited framework for the effective design of UIs is Molich and Nielson's (1990) usability heuristics. Despite major advancements in technology in the thirty years since their inception, these heuristics remain the most commonly used frameworks by software designers and developers. These 10 heuristics (Figure 4; overleaf) are used to provide general guidelines to software developers to ensure that the user experience while using a digital platform or interface allows for intuitive interactions and easily executable tasks. For example, the first heuristic in this framework emphasises the importance of a visible system status. Adhering to the principles of this heuristic requires that the users are given feedback on what is happening in the system within a reasonable timescale. Apple (2018) encourage their software designers to use an auditory cue such as a 'beep' or a 'whoosh' to indicate that a task is complete.

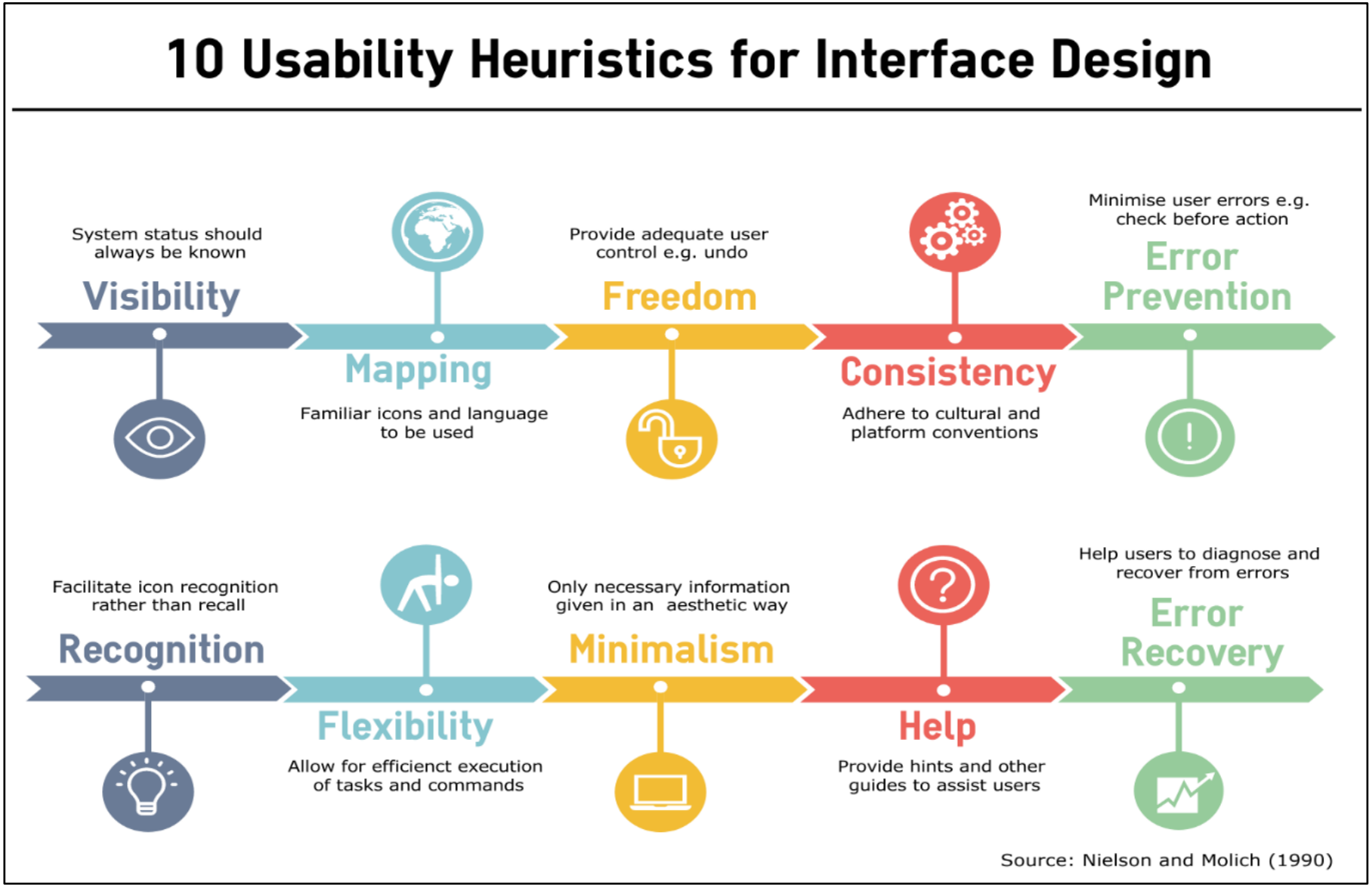


Figure 4 Usability Heuristics for Interface Design (Molich & Nielson, 1990)

Adherence to these heuristics is associated with improved UI design as demonstrated. However, to maximise usability, Harms and Adams (2008, p. 4) assert that each component of a UI, regardless of the prospective device, industry or purpose, must be also designed 'with consideration of the knowledge, expectations, information requirements, and cognitive capabilities of all possible end users'. Therefore, the interface design of the CBE for CS should take into consideration the specific needs of students in an online environment. While literature on the design and evaluation of commercial or industry specific software is readily available from companies such as Apple (2018) and Microsoft (2018), studies on the UI design and development for online assessments or assessment software are limited. At present, only one study by Dembitzer, Zelikovite and Kettler (2017) was found to explore the development of a UI system for a literacy based CBE with 131 high school students. In this study, a partnership between psychologists and software designers was established to develop appropriate design procedures and processes for this CBE to determine how the psychometric concerns of accessibility, reliability and validity can be balanced with issues of usability. While this study took great care with the UI design for the planned CBE, it is unclear if other CBEs are approached in the same manner. This is worrying as what research is present indicates that certain aspects of a UI's design need to be carefully constructed to ensure usability in a test-taking situation. In particular, issues surrounding the navigation, recommended user interface tools and the scoring of test questions should be considered.

5.2 Navigation

Navigation relates to how the examinee moves around in a test. According to Luecht and Sireci (2011), there are two aspects to navigation in CBEs: the visual style and freedom of the navigation control and the test-taker's ability to review previously seen questions or submitted responses. In relation to the visual style of the navigation control, this can vary significantly between CBE delivery systems. Some CBEs merely use a 'forward/next' and 'back' controls to allow the examinee to navigate item-by-item. Luecht and Sireci (2011, p. 12) note that other CBEs allow the test-taker to 'jump' between items using a full-page 'review screen' to 'display all of the items, the examinee's item response (if any), and any flags left by the examinee indicating that he or she wants to possibly review the item later'.

Surveys of test-takers and students have reported that they prefer being able to freely navigate through an examination, rather than being forced to answer a single item at a time, with no opportunity to 'skip ahead' or 'look back' (e.g. Parshall, Spray, Kalohn, & Davey, 2002; Walker & Handley, 2016). Certainly, this is consistent with the 'visibility' and 'freedom' principles of Molich and Nielson's (1990) usability heuristics. Qualitative responses from undergraduate students participating in Walker and Handley's (2016) research considered easy navigation to be essential to the usability of an online examination. They asserted that free movement of students in a section of a CBE was necessary to accommodate students' differing preferences for the sequencing of questions, a common test-taking strategy in PBEs. Therefore, in order to facilitate students in their use of test-taking strategies, students should be able to see all the questions on a single screen at the start of the test and whenever they deem necessary. This recommendation can be found in previous studies including Frein (2011).

Item review, that is, whether the test-taker or student can go back to an answered or unanswered item and change their answer, can be prohibited in a CBE if the interface prevents backward navigation. Navigation between items has no such restrictions in a PBE. Vispoel's (2000) research demonstrated that students have a strong preference for item review opportunities, especially those students who are prone to test anxiety. This was also found in Walker and Handley's (2016) research. Yet, it should be noted that the option to skip, review, and change previous responses in a CBE has not been shown to consistently influence students' ***scores*** (Hadadi, Luecht, Swanson, & Case, 1998). When Hardcastle, Herrmann-Abell and DeBoer (2017) explored whether the restrictions on item reviews influenced elementary, middle and high school students' test scores in a CBE (n=34,068), they found that scores for elementary and middle school aged test-takers were negatively affected by a CBE interface that did not allow them to return to questions they had previously completed or skipped. This finding did ***not*** apply to high school students who would be most similar in age to the test-takers in the upcoming CS exam. Based on this research, it appears that older students do not need item review capabilities in their CBEs as much as younger students. However, test-taker *preference* for the ability to review items has been consistently found in research (e.g. Walker & Handley, 2016). Therefore, while it is not essential that students be able to review their work according

to research, it appears that student satisfaction with the CBE interface may be enhanced by such a feature.

5.3 User Interface Tools

As the implementation of CBEs has become more widespread, state testing authorities in the US have now begun to embed various test-taking tools in the interfaces of these CBES (DePascale et al., 2016). Many UI tools are available but the most common ones found in CBEs include progress bars, clock or timers, word processing features and the provision of annotation tools. Table 2 (overleaf) provides an overview of the features associated with such tools.

Table 2 Overview of common UI tools used in CBEs

Tool	Variations	Purpose
Progress Bar	- Progress information on TBA - Progress information on a task/ TBA section	Progress indicators provide information to test-takers about task duration.
Countdown Clock	- Time elapsed/left - Time	Timers and clocks allow test-takers to manage their time effectively.
Word Processing	- Spelling and Grammar Check - Text formatting	Allows for the creation and formatting of documents
Annotation Tools	- Highlighting/ Marking - Comment boxes	Highlighting key words/questions and recording notes facilitates essential test-taking strategies.

In 2003, Fulcher noted that there was very little published literature on the application and value of UI tools for CBEs despite rapid growth in the industry. Unfortunately, little has changed in the intervening 15 years. While some UI tools like word processing and annotation tools has received some attention in assessment research, the majority have not been research in great detail. Further information on the possible value of some of these UI tools in a CBE can only be hypothesised from research

conducted in non-assessment based contexts. Findings from such research will now be discussed in relation to key UI tools that could be included in the upcoming CS exam including progress bars, countdown clocks, word processing tools and annotation tools.

5.3.1 Progress Bars

Progress bars are used to show the current state of a particular task or activity that the system or user is currently engaged in. Myers' (1985) original study on the subject found that people prefer to have progress indicators in online environments. Certain psychological phenomena are associated with this preference such as the Gestaltian Law of Closure (Wertheimer, 1923; humans have a need to complete items) and the sunk cost effect (Arkes and Blumer, 1985; once a visible amount of time or effort has been committed, individuals tend to persist with a task regardless of logic). Villar, Caellegaro and Yang (2013) noted in their meta-analysis of the subject that progress bars are a common feature of most UIs and can come in many forms as seen in Figure 5.

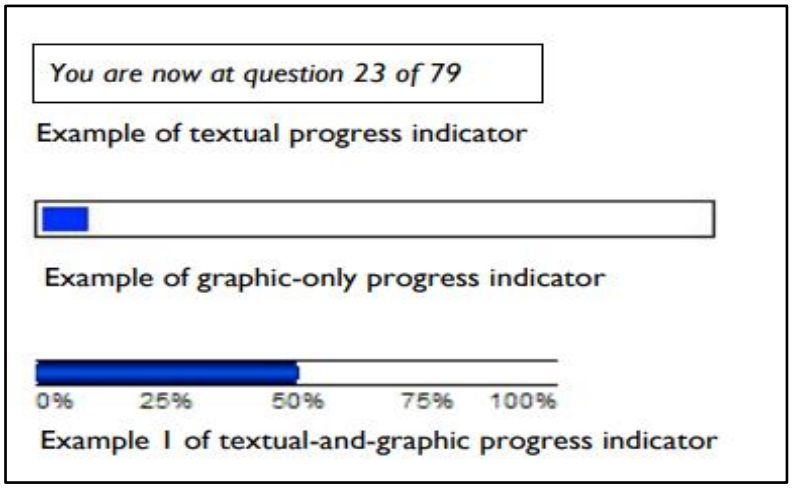


Figure 5 Sample Progress Bars (adapted from Villar et al., 2013)

In line with Nielson and Molich's (1990) usability heuristics of visibility and freedom, Conn (1995) asserts that progress indicators should specify, amongst others, the acceptance, scope and overall progress of a task to the user. Unfortunately, the actual impact progress bars have on test-taker performance and behaviour is less clear. No research directly related to the use of progress indicators in CBEs is currently available. Instead, research on the efficacy and use of progress indicators in online environments

have been primarily conducted by survey researchers. Although this research provides some useful insights into the role of progress bars on human behaviours in online environments, the findings may not be directly transferable to online assessment situations. Surveys are optional for participants and it is their choice whether or not they complete the task. CBEs on the other hand require student to complete the task regardless of their desire to 'drop out'. Therefore, some caution should be exercised when attempting to relate the following research findings to student performance in CBEs. However, they may provide some insight into their value to students when completing online tasks.

A between groups study (*n*=3179) conducted by Conrad, Couper, Tourangeau and Peytchev's (2010) explored the effect of three types of indicators. Always-on, Intermittent (at nine transition points in the questionnaire) and On-Demand (users could obtain feedback by clicking a link labelled 'Show Progress'). Based on their research findings, the authors recommended that an intermittent feedback schedule should be provided if progress indicators are used in online environments. In relation to CBEs, this type of feedback could be provided between section breaks or when the student passes certain milestones. Progress feedback in the forms of clocks and timers should also be considered by users in CBEs.

5.3.2 Clocks and Timers

In online environments and CBEs, clocks and countdown timers aim to provide the user with temporal data. As noted by Lim, Ayesh, Stacey and Tan (2017, p. 99) time information in exams is essential for students as it allows them to generate the 'optimal plan and strategies to complete the test based on the time resources provided'. Despite their role in facilitating the implementation of effective test-taking strategies, the ideal way to present this information to students has not been extensively researched or even discussed in relevant literature. Temporal information could be displayed using clocks that tell the current time or through the use of timers that will display how much time has elapsed.

In an exploratory study involving 119 undergraduate students selecting different UI tools to personalise an online assessment to their own needs and preferences, Karim

and Shukur (2016) presented four options for displaying countdown timers and clocks: descending countdown timer (time elapsed), ascending counter (time left), a traditional clock (analogue) or no clock at all. According to this research, test-takers do desire timing information in CBEs as only 1% of the sample chose to display no form of timer. A descending or elapsed timer was the most popular choice (42%) followed by a standard analogue clock (31%) and an ascending timers (26%).

It is important to note that the preferences of this particular group may not be representative of all potential students. Furthermore, this study explored *preferred* online examination design and UI tool usage. It did not explore the influence of such clocks and timers on student *performanc*e. Allowing students to choose what type of tool they would prefer does not guarantee that tool's efficacy in maximising their performance. Regardless, access to some form of timing tool appears to be desirable to test-takers in a CBE. Interestingly, a study conducted by Lim et al. (2017) revealed that countdown timers increased self-reported stress and error levels among students completing an online test of arithmetic. A standard analogue clock reduced reported stress levels. The results of this study suggest that clocks rather than timers appear to a more effective way to display temporal information to test-takers. These results should be interpreted with caution as Lim et al.'s (2017) study contained several methodological anomalies and irregularities in relation to statistical analysis but it does provide some insight into the possible effects of timing devices in testing scenarios.

In relation to the CS exam in 2020, careful consideration should be given to timing issues. For example, if a student device's built-in analogue clock is used, clocks between devices should all be synced with each other and any external analogue clocks. This will ensure standardisation. Similarly, students should all start the exam at the same time so that students do not receive a timing advantage. Furthermore, if a student experiences a full or part technology failure during the exam, procedures should be in place to ensure that all students are afforded the same amount of time to complete their exam.

5.3.3 Annotation Tools

Annotation refers to marks made by readers on reading documents or on a companion document often called 'scratch paper' (Prisacari & Danielson, 2017). Marshall

(1997) suggested that two forms of annotations are used by readers – explicit (such as text-based or graphical notes) and inexplicit (such as highlight, underline, asterisk, arrow, and graphics). Marshall (1997) notes that inexplicit annotations by readers are usually employed for procedural reasons as they are often used to signal to the reader what is or is not known or whether or not something will require a review at a later time. Explicit annotations usually take the form of notes that record thoughts or ideas or to solve a problem.

Providing students the option to 'mark' a test item with a particular symbol like question mark, as 454 9th grade Taiwanese students did in Chen, Ho and Yen's (2010) study did, resulted in improved English test scores for some students. Marking test items did not have a positive impact on students with high or low English ability but did for those with moderate English language skills. Rogers and Bateson's (1991) discussion on this issue could explain this result. Identifying clues embedded in the test items is a key test-taking strategy. Marking such clues by underlining or marking allows test-takers to concentrate on these clues with increased ease. In Chen et al.'s (2010) study, high ability students did not seem to benefit from marking because the required knowledge was known to them – they did not need to search for clues. Marking did not approve the performance of low ability students as they did not have the knowledge to answer correctly even if they did use appropriate marks. In contrast, marking test items appeared to have encouraged average ability students to focus on the information that would allow them to recall the information needed to select the correct answer.

Therefore, it seems that providing test-takers with annotation tools in CBEs would be beneficial**.** Furthermore, Jackel (2014, p. 74) acknowledges that annotations are 'particularly important to certain kinds of thinking'. Studies of annotation patterns presented at various conferences by Marshall (1998) and Plimmer and Appleby (2007) show that science text books are more likely to have notes pencilled in the margins than any other textbooks. Plimmer and Appleby (2007) note that these annotations were used to solve problems and record ideas and interpretations. However, it can be difficult to engage in such note-taking strategies in CBEs as the standard keyboard-mouse interaction paradigm does not always lend itself to the unconscious note taking behaviours associated with paper based assessments. This is why scratch paper is often provided when CBEs are deployed. Research by Prisacari and Danielson (2017) found

that 81% of students (*n*=221) undertaking a CBE to assess their knowledge of an undergraduate chemistry course used the scratch paper to answer three or more algorithmic questions. In contrast, 95% of students did not use the scratch paper for definition questions. Given the range of algorithmic problems and deployment of code-tracing associated with CS, it may be best to provide CS students with scratch paper in 2020 alongside the CBE. However, it is important that students do not have to 'replicate' their work on scratch paper in the CBE as this can be particularly frustrating for students (New Zealand Qualification Authority, 2014). If the question in the CBE requires that they 'show their work', students should be able to conduct their 'rough work' within the CBE.

5.3.4 Word Processing Tools

The upcoming CS exam may require students to compose some discursive responses like short-answer questions. If this is to be done in a computer-based environment, then word-processing tools or software will need to be made available to the students. In line with Nielson and Molich's (1990) 'mapping' heuristic, the word processing tools and software used in a CBE should be familiar to the students and closely resemble the functionality associated with commercially available word processing software. It should also be acknowledged that the process of composing text using a keyboard is different to using a pen and paper. When handwriting answers, the student's success is highly dependent on their ability to plan out their writing in advance. In contrast, using word processing software and tools (e.g. spellcheck, text formatting etc.,) makes editing a text much easier. This can somewhat diminish the role of planning thus exemplifying the key differences between handwritten and typed answers. Preliminary research has indicated that these different approaches can influence student performance in exams.

In their series of meta-analyses on the use of formative assessments for writing Graham, Harris and Hebert (2011) found that students writing on a computer scored higher than students writing with paper and pencil. In the seven true and quasi-experiments that were consulted in this meta-analysis, where one group of students wrote using word processing during testing and the other group wrote by hand, a statistically significant, moderate effect of 0.54 in favour of a word processing approach

was obtained. It appears that composing handwritten answers may underestimate a student's ability to effectively convey thoughts and ideas. However, this finding may not apply to all students. Graham et al. (2011) also found that students with little experience using a computer to write responses had higher scores when handwriting essays. This finding was also reported from the US (National Centre for Educational Statistics, 2012).

Therefore, if students are not proficient at writing responses on computers, they will receive a better score if they handwrite their answers. If they are adept in the use of word processing software and writing text on computers, students will receive a higher grade through this medium. However, Graham et al. (2011) caution that student performance can be further moderated by the marking behaviours of examiners.

5.4 Scoring Procedures

It must be remembered that test-takers will not be the only individuals interacting with the CBE. Examiners will be assigning scores to student responses within the CBE as well. In relation to the marking of items in CBEs, the automatic scoring of test responses is a major advantage of CBEs over traditional PBEs. However, some items are easier to score. For example, multiple choice, figural response (those that involve manipulating images etc.,) and short answer questions are relatively easy and efficient to score as computers can score such items automatically. In contrast, constructed-response items, like short-answer questions, are more problematic. O'Leary, Scully, Karakolidis and Pitsia (2018, p. 4) note that automated text scoring systems are at a stage of development where they can be considered 'efficient, impartial and objective' but that their ability to 'understand the nuances of human communication may still be some time away'. Therefore, discursive questions completed on computers will need human scorers.

If discursive questions (e.g. short answer, essays) are to be included in the CS exam, then it is important to consider what impact a typed response would have on examiner markings and ratings. In a review of five studies comparing identical typed and handwritten essays, Graham et al. (2011) found that teachers give statistically lower scores to a printed computer copy of an essay than to an exact handwritten copy of the same paper. Russell and Tao (2004) studied how response format affected raters' judgments for middle and high school students. Essays were presented to raters in one of

three formats: handwritten; printed in single-spaced 12-point text; and printed in double-spaced 14-point text. The handwritten versions were found to receive significantly higher scores than the typed versions, but no differences were found between the two forms of typed essays. Results also indicated that the raters detected more spelling and syntax errors than when essays were presented in print. Findings from such research will need to be considered in relation to the upcoming CS exam.

5.5 Guidelines for CS Examination: Human-Computer Interaction

- The design of online environments has been informed by Nielson and Molich's (1990) usability heuristics for a number of years. The CS examination should also be designed according to these principles.
- According to research, students in CBEs prefer to have on-demand access to an overview of the test items in an examination to facilitate test-taking strategies. The facility to 'jump' between questions is also preferred among students. These preferences should be considered when designing the CS CBE.
- Although there is no evidence that it influences performance in teenage test-takers, the ability to review items before final submission is highly desirable among test-takers. Students should be able to review test items within the CBE before submission.
- If progress bars are to be included in the CBE for CS, evidence from survey research suggests that it should be provided using an intermittent display schedule e.g. after every section etc.
- It may be best to allow students to choose what *type* of timing device they prefer within their CBE (countdown timer etc.,) or use a standard clock that Leaving Certificate test-takers will be more familiar with.
- Clear guidelines regarding timing procedures should be given to schools and test centres to ensure standardisation.
- The provision of offline and online annotation tools is recommended for CBEs in order to facilitate essential test-taking strategies.
- Any word processing tools that are included in the CS CBE should closely resemble the functionality associated with commercially available software.

- If discursive or text-based short answers are to be completed within the CBE, efforts must be put in place to ensure that prospective test-takers have the required experience and proficiency using word processing software.
- Examiners for the CS exam in 2020 may need to be trained to avoid scoring biases that may be associated with response format. Examiners will need to be provided with training on the biasing effect of typed and handwritten responses.

6.0 Test Design

Evaluating students' knowledge and learning in CS is challenging (Kallia, 2018). The design of appropriate assessment instruments that can identify a student's proficiency and attainment of curricular content is difficult to achieve in CS courses given the subject's complexity and the array of skills involved. To overcome this, Bloom's updated Taxonomy (Anderson et al., 2001) is frequently used to assist in the development of CS assessment tasks and exam questions to ensure that a range of lower and higher order thinking skills are being elicited from test-takers, specifically: Remembering, Understanding, Applying, Analysing, Evaluating, and Creating. Work by Lopez, Whalley, Robbins and Lister (2008) indicates that the content of CS exams for introductory courses should include a variety of tasks that incorporate these skills to create a 'hierarchy' of programming skills. Activities would include code reading, code writing and code-tracing along with standard questions requiring students to recall and explain key information or concepts. Questions that assess such content can take many forms and can include multiple-choice, figural response and constructed response items. Thanks to recent advancements in technology, constructed response questions that would allow test-takers to write and develop code in an authentic environment (also called a *sandbox*) are now possible. Yet, the design of all these questions for a CBE should be carefully constructed, taking into account best practice guidelines from the field of education and testing as well as previously devised CS exams from third-level institutions. The organisation of these items should also be closely considered.

6.1 Multiple Choice Questions

Traditional selected response items include a text or image based prompt as a stimulus followed by a range of possible responses from which the test-takers must select the best choice(s). Selected response items can vary in type and complexity but are characterised by the test-taker selecting a response option. These items can include true/false questions, extended matching questions, multiple choice questions and situational judgement items. The most commonly used selected response items in Leaving Certificate exams are multiple-choice questions (MCQs). Testing literature asserts that items that MCQs have many advantages including 'efficient administration, automated scoring, broad content coverage and high reliability' (Wan & Henly, 2012 p. 59). However, when Shuhidan et al. (2005) explored CS instructors' perspectives of MCQs, a negative perception of these types of items was found. The instructors in this study felt that MCQs did not provide an appropriate measure of CS students' abilities and tested only lower order thinking. Yet, Woodford and Bancroft's (2005) guidelines for educators in Information Technology courses assert that effective MCQs that can assess higher levels of knowledge can be constructed. Scully (2017) supports this argument and gives clear strategies that outline how this can be done. These have been outlined in Table 3 (overleaf).

Table 3 Strategies for authoring MCQs that assess higher-order thinking skills (based on Scully, 2017)

Strategy	**Explanation**	**Example**
Verb Selection	Verbs from higher levels in Bloom's Taxonomy should be converted to their noun derivative and paired with a knowledge level verb (Dickenson, 2001). For example, 'describe' (comprehension level verb) could be used to create this question stem: 'select the best description'.	Identify the most appropriate modification from the list below to allow this code to repeat 8 times instead of 4.
High-Quality Distractors	Distractors that are superficially similar to the targeted concept demand a high level of critical thinking and judgement. MCQs which ask students to select the 'best' answer, where all of the given options are theoretically plausible can effectively assess higher order thinking skills.	Which best describes a fragmented hard drive? A. Data files are stored incorrectly. B. Data files are corrupted. C. Data files are not stored in consecutive clusters.
Item Flipping	Traditional MCQs name the concept being assessed and the asks the test-taker to select the most appropriate definition (e.g. What is a 'loop'?). 'Flipped' MCQs require test-takers to identify the underlying rule or concept presented in the question stem.	The above code represents which of the following? A. A sentinel loop B. A while loop C. A do… while loop D. A conditional loop
Conceptual Connections	MCQs that assess higher-order thinking skills should 'tap' multiple neurons (Scully, 2017). This means that MCQs should require students to demonstrate their knowledge of multiple ideas or concepts and understand the relationships between them	Using the **`goto`** inside **`for`** loop is equivalent to: A. `continue` B. `break` C. `return` D. none of the above

Therefore, MCQs, are an appropriate item to include in a CS exam. They can easily measure lower-order thinking skills and, if constructed properly, can also assess higher-order thinking skills. However, CBEs can allow MCQs to represent authentic scenarios to students using multimedia prompts. In CBEs, prompts can take many forms such as text, images or animations. It is important to understand what impact these different prompts have on test-taker performance. For example, text-based prompts will likely be the dominant form of stimulus used in MCQs for CS exams given the text-based nature of coding. Students may be asked to identify from a list of options what a code does or they may be required to match a code to a particular definition, Best practice guidelines such should guide the construction of MCQs that target higher-order thinking skills. However, images of computer parts or arrays may also be included as stimulus prompts for the upcoming CS exam. Reasons for the deployment MCQs with these types of prompts should be monitored to ensure that appropriate judgements are made.

In a small-scale, mixed-methods study conducted by Vorstenbosch et al. (2014), seventeen first-year medical students answered Extended Matching Questions (EMQs; a version of an MCQ) regarding their understanding of anatomy, using either labelled images or answer lists in a paper-and-pencil test. Vorstenbosch et al. (2014, p. 107) found that EMQs with and without images seemed to 'measure different skills, making them valid for different testing purposes'. Students used more cues from EMQs with images and visualised more often in EMQs with text-based answer lists. Items without images seemed to test the quality of students' mental images while questions with images tested their ability to interpret visual information. Therefore, both response formats should be used to construct MCQs in CBEs to offer a broader representation of the targeted skill being assessed.

Other multimedia prompts, like animations, are also possible in CBEs. *BlueBook* is a secure, reliable, cross-platform desktop application for administering CBEs that has been used in Stanford University for a number of years (Piech & Gregg, 2018). This platform (which will be discussed in more detail in the ***Sandbox*** part of this section) 'supports animations and playable demos of questions so that students can understand the task better' (Malik, 2018). This illustrates how animations are becoming more common in CBEs and will likely be employed in CS CBEs. In an experiment comparing the performance of novice and expert drivers (n=100) in a CBE involving multimedia stimuli,

Malone and Brünken (2013) found an interaction effect between stimulus type (animations, images) and expertise level. Novice drivers benefitted from the higher ecological validity of the animated presentation as they did not have to infer relationships and motion from a static image. While this facilitation effect could cause some concerns in other exams as it may result in the incorrect grading of test-takers, this could prove useful in introductory CS exams. As mentioned previously, programming in CS courses is a complex skill that takes many years to achieve expertise. Exams for novices are therefore often hard to devise and it can be difficult to provide all the support novice programmers require in a test-taking setting. Using animations as prompts for MCQs may be one way of providing support to novice programmers in examinations.

Scully (2017) admits that MCQs have certain limitations. However, if carefully constructed to include both higher and lower order thinking skills as well as a variety of stimulus prompts, they can be a valuable component to any CS exam. It is important to remember that they should also be complemented by other item types in CBEs.

6.2 Figural Response Items

Figural response items depend on material such as illustrations, graphs and diagrams. They require examinees 'to manipulate the graphic elements of an item, click on one or multiple 'hotspots' on an illustration, or complete a diagram by dragging and dropping elements' (Wan & Henly, 2012, p. 63; Figure 5). As with MCQ items, figural response questions require the test-taker to 'select' their answer. However, figural response items, depending on their construction, are distinguished by the increased level of freedom a test-taker has in *what* they select and in what they *do* with the item selected. An early study by Martinez (1991) determined that figural response type items in a paper-based assessment of science for 4th, 8th and 12th grade students were found to be more effective at discriminating between students with differing proficiency levels than select-only, multiple choice items (Figure 6; overleaf).

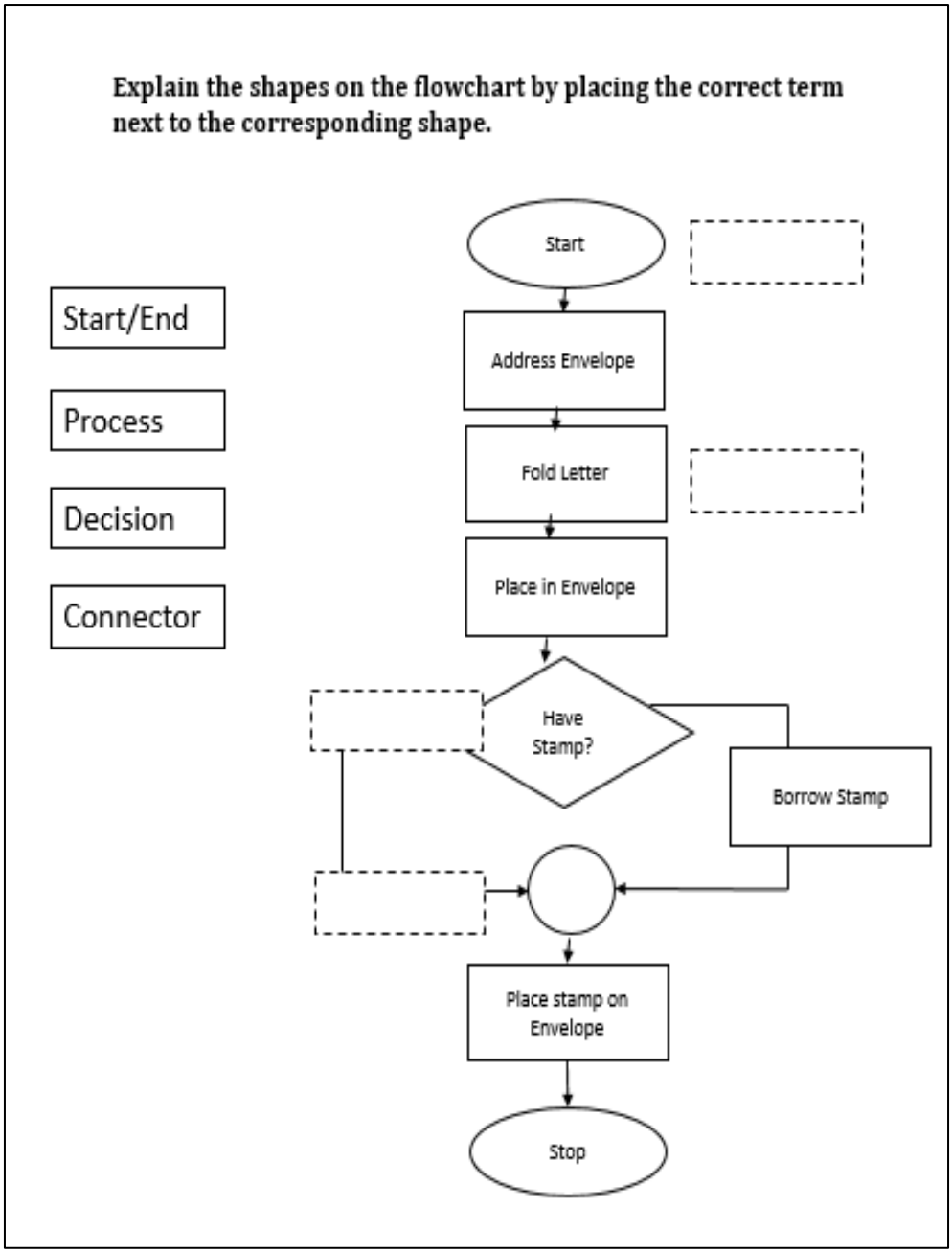


Figure 6 Example of a Figural Response item (drag and drop)

Wan and Henly (2012) studied the reliability and efficiency of different item types in CBEs including figural response items that employed the 'drag and drop' and 'hotspots' functions in a CBE interface for school aged students in a state-wide science based achievement test. Figural response items were found to be as good as MCQs items in providing information about student ability. A more recent study by Woo et al. (2014) on

behalf of the National Council of the State Boards of Nursing found that rates of random guessing also appeared to be reduced when responding to figural response items when compared to MCQs. However, students required significantly more time to answer items involving drag-and-drop response actions. It should also be noted that drag-and-drop and hotspot items should be carefully designed to ensure that they do not provide test-takers with unintentional cues that would guide their selection of answers e.g. target items 'snapping' to a particular location if it is correct etc., This would introduce construct-irrelevant variance into the testing scenario which should be avoided at all costs.

Figural response items have received only limited attention in research and their empirical value is difficult to determine. However, it is likely that these items can provide richer diagnostic information than MCQs. While it is possible to infer from the incorrect option selected in an MCQ what student's difficulties and misconceptions are, it is easier to record, analyse and assess cognitive processes and problem-solving strategies in figural response items as the student must 'do' something with the response option that has been selected or identify for themselves what must be selected. Therefore, creating these times of items for the upcoming CS exam could prove useful. This function of figural items could be quite useful for formative assessment. While the Leaving Certificate is summative in nature, aggregate data related to student performance on these items could be used to inform general teaching and learning in CS. If these items are used in mock exams, they could provide teachers with useful data regarding individual students' strengths and weaknesses.

Certain types of CS concepts and skills would lend themselves well to this particular type of question. For example, Lopez et al.'s (2008) PBE for undergraduate computer programmers had a variety of tasks that could be easily adapted to a figural response item in a CBE. In the 'basic' types of questions, Lopez et al. (2008) required students to select a piece of code and match it to the correct term in a given list. Another question asked them to identify syntax errors within a piece of code. A Parson's Puzzle would also be suited to this form of question. A Parson's Puzzle (Parsons & Haden, 2006) requires students to take a set of lines of code, presented in random order, and place those lines into the correct order to perform a given function (Figure 7; overleaf). According to Lopez et al. (2008, p. 4), these puzzles 'require students to apply their knowledge of the common patterns in basic algorithms, apply some heuristics (e.g.

initialize a variable before using it) and perhaps also, to a degree, manifest some design skill'. Such items could be deployed as MCQ or as a figural response item in a CBE. Denny et al. (2008) found that Parson's Problems effectively evaluated CS skills associated with more difficult code writing tasks but using an approach that was easier and better suited to CS students whose skills were developing at a slower rate. They were also considered particularly useful for the identification of student misconceptions.

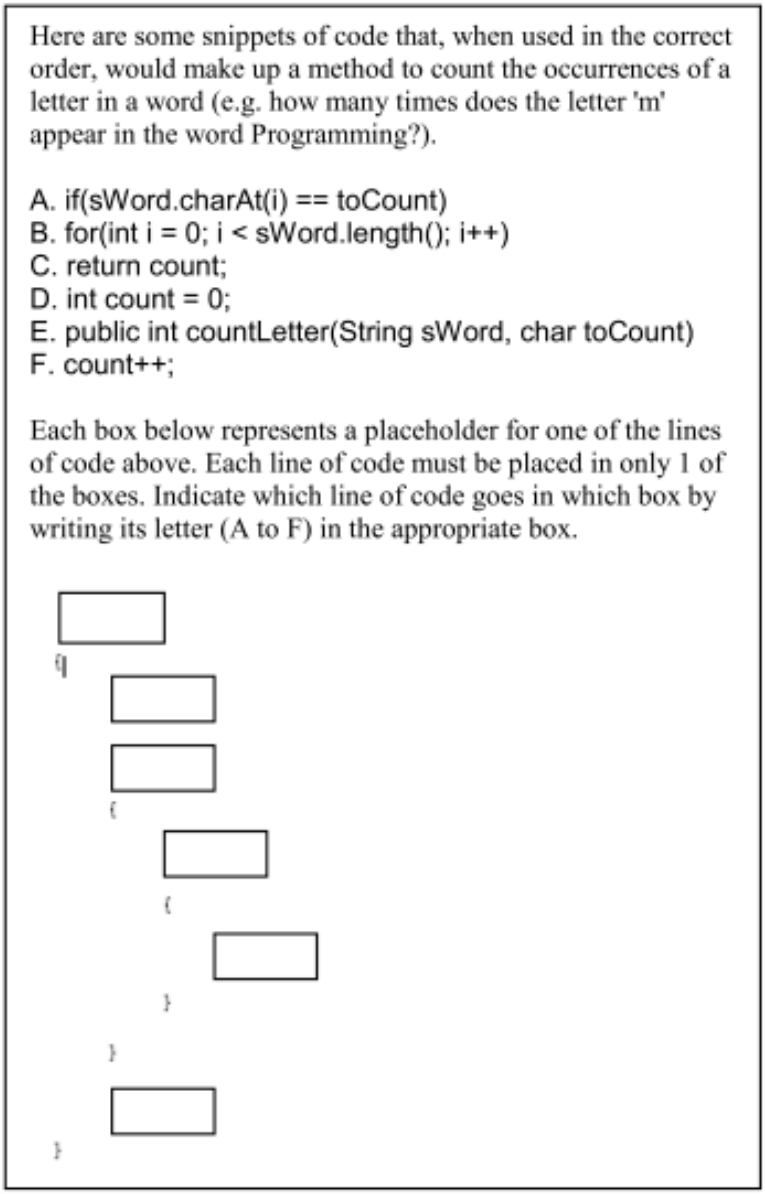


Figure 7 Parson's Puzzle from a PBE (Lopez et al., 2008)

6.3 Constructed Response Items

Wan and Henly (2012, p. 63) define constructed response items 'as those which require students to create an alphanumeric response which can vary in length'. These include short or extended written responses (e.g. fill-in-the-blanks, essays, spreadsheets). Although this item type is considered to be very traditional, it is still used extensively in CBEs. In the case of CS, writing code is the most common form of constructed response question. When they are asked in PBEs, students must handwrite long lines of code in essay form. However, unlike PBEs, CBEs can provide programming environments ('*sandbox*') where students can write and develop code using the keyboard. A sandbox is a virtual testing environment where programmers can securely run and develop new and untested pieces of code without risk of contaminating other previously designed software. While a sandbox can be easily included in a CBE, what features should be included should be take into consideration

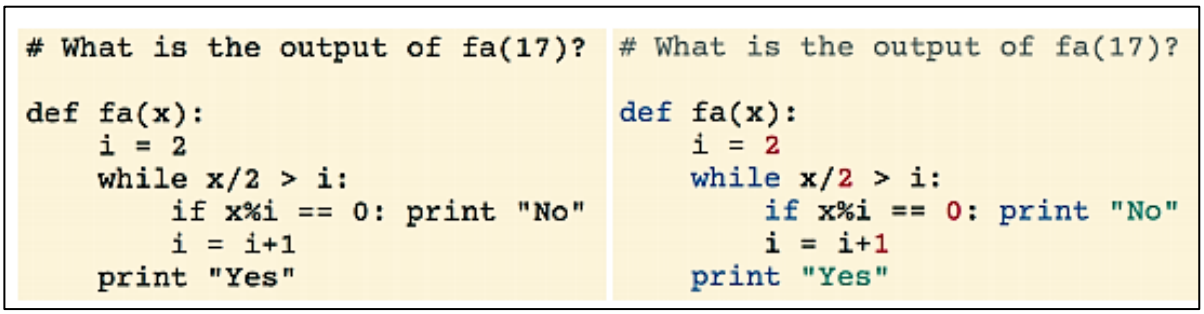
6.3.1 Sandboxes for CBEs

Sandboxes for novice CS programmers usually bear some resemblance to an Integrated Development Environment (IDE) for a particular programming language (e.g. Python, Java) but with some modifications. An IDE is a visual environment that assists programmers through the deployment of a variety of features such as auto-indentation, line numbering, syntax highlighting, auto-completion and drag-and-drop coding (Dillon, Anderson & Brown, 2011). IDEs can debug, compile and execute code as well. In a classroom situation, it allows the students to both see what was erroneous in their initial solution, as well as give them the opportunity to fix these errors by themselves. In relation to the CS exam, what of the afore-mentioned IDE features should be enabled and disabled in the proposed sandbox for *assessment* purposes should be informed by relevant research.

In particular, the compiling and execution of code in a CBE for CS has received some attention in the limited amount of literature available. Haghighi et al. (2005) evaluated students' attitudes (n=72) towards computerised CS exams. Although the students found that the most useful aspect of the exam was access to the compiling and debugging facilities, these features also caused students a significant amount of anxiety. If the students realised that their code did not compile properly in the exam, students

reported feeling higher levels of stress and anxiety. This supports anecdotal evidence cited in Piech and Gregg's (2018) research. When the authors first approached the CS department at Stanford University about their hopes to trial a computerised CS exam, there was some resistance as a previous trial of a computerised CS exam had been unsuccessful. In this trial, students had the ability to compile their code and then run it through a test suite. This proved 'disastrous' (Piech & Gregg, 2018, p. 4) as students spent too much time trying to fix their code. Students did poorly in the exam as they often did not finish the test paper because of their tendency to 'fixate' on these questions.

Only a restricted amount of research exists in this area and some of it is anecdotal. However, it does appear that students should not have the option to compile and run code in exam situations. From a pedagogical standpoint, Piech and Gregg (2018, p. 4) acknowledge that disallowing the compiling or running of code ensures that students do not simply try 'different solutions until they finally land on one that works'. The authors believe that including these features in an exam would require providing other supports (e.g. giving answer code length descriptions) to students to help them manage their time effectively. In contrast, other features of an IDE have already proved useful in a sandbox for students in exam situations. Syntax highlighting, the 'meaningful colouring of text in coding environments' (Sarkar, 2015), was reported to increase readability and comprehension speed for novice programmers. In Sarkar's (2015) small-scale exploratory study (n=10), participants were required to mentally compute the output of a Python function for a given set of arguments for highlighted and non-highlighted code (Figure 8). Using eye-tracking technology, syntax highlighting was found to incur a lower level of cognitive load thus ensuring that the test-taker could focus on the key points of the code. However, Öqvist and Nouri (2018) acknowledge that the colour coding should be one that the student is familiar with if these results should be replicated.



```
# What is the output of fa(17)?

def fa(x):
    i = 2
    while x/2 > i:
        if x%i == 0: print "No"
        i = i+1
    print "Yes"
```

```
# What is the output of fa(17)?

def fa(x):
    i = 2
    while x/2 > i:
        if x%i == 0: print "No"
        i = i+1
    print "Yes"
```

Figure 8 Code with and without syntax highlighting (from Sarkar, 2015)

Taking into account this research, and being unable to find an IDE or editor that took this research into consideration, Öqvist and Nouri (2018) designed a test platform for their study exploring the comparability of assessment mode in CS exams. The test environment was a text editor rather than an IDE as it did not allow students to compile or test their solutions. Instead, this editor was designed to support the editing, structuring and readability of code. The environment supported auto-indentation, line-numbering and syntax highlighting. The students could write, delete and move code. Test-takers in this study were particularly appreciative of their ability to edit their code as it allowed the students to engage in an iterative problem-solving approach that was consistent with what they had been taught to do previously.

BlueBook is an open source cross-platform Java program for administering CS exams that is currently in use in CS undergraduate courses at Stanford University in the US (Malik, 2018). The program runs in full screen mode and does not allow students to switch between programmes. In line with HCI literature, problems and student solutions occupy the same screen as demonstrated in Figure 9.

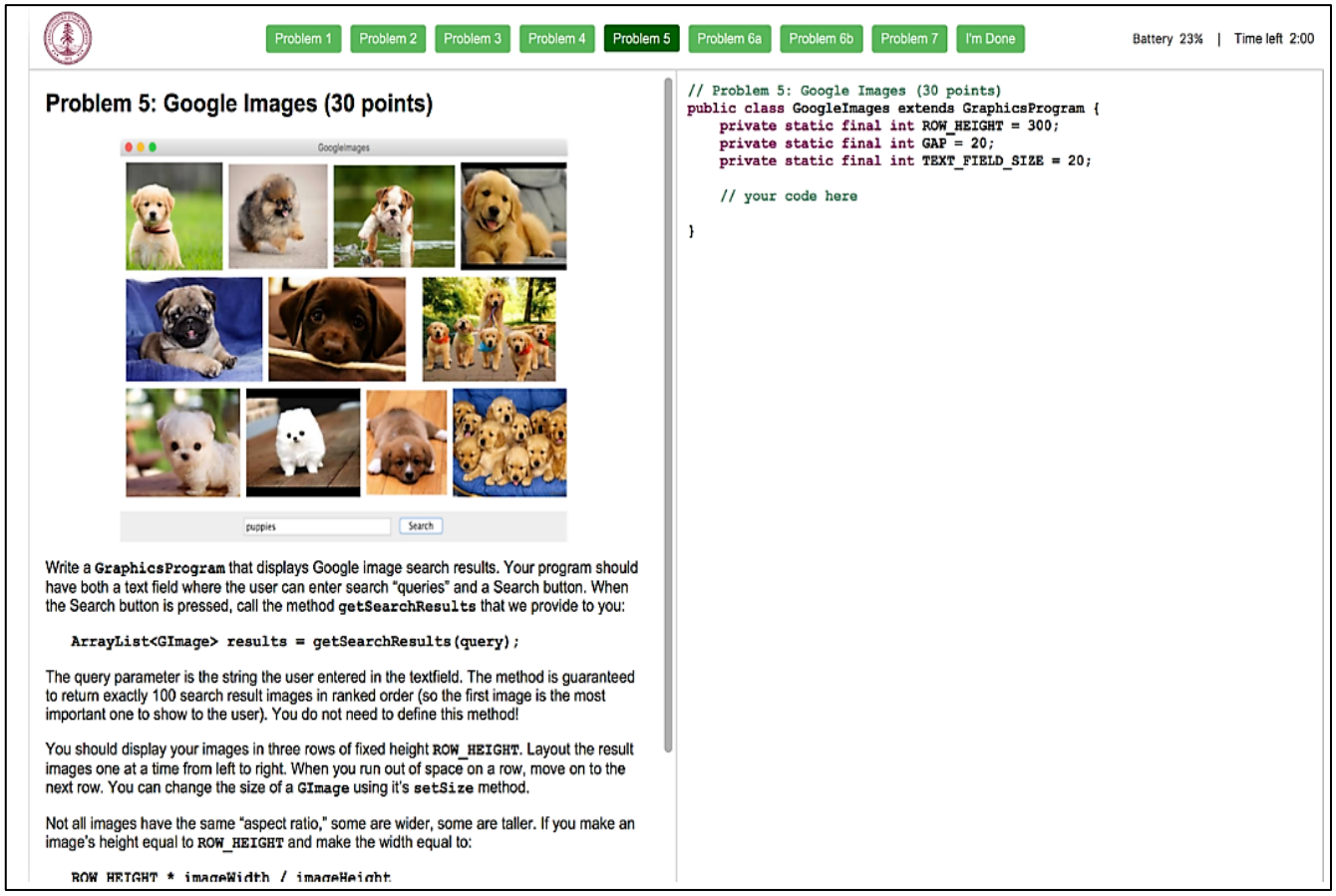


Figure 9 BlueBook Screenshot (from Piech & Gregg, 2018)

Key features of the *BlueBook* examination environment include:

- Code answers can be syntax-highlighted based on programming language.
- Students cannot compile or execute code.
- UI tools can be included e.g. a count-down timer. This can be modified for students who need extra time or who prefer different types of clocks.
- *BlueBook* keeps the computer screen maximised and does not allow students to switch to other programs.
- *BlueBook* supports animations and playable demos of questions so that students can understand the task better.

Feedback from students who used *BlueBook* for their summative assessments was largely positive (Piech & Gregg, 2018). Interestingly, students reported that the timer was particularly beneficial to have to refer to as they were developing their code. In response to this, Piech and Gregg (2018) are considering an individual 'timer per problem' that will show students how much time they have spent on a particular problem. Alternatively, an 'optimum response time' could be allocated to each question. Delen (2015) explored this in a CBE for geography teachers in Turkey. The displayed questions in this CBE had optimum response times. Students were not obliged to 'move on' once the time had elapsed - they could spend more or less time at the question. Delen (2015, p. 1468) found that test-takers spent 'a sufficient and reasonable amount of time when answering the questions' when an optimum response time was displayed compared to those who did not see an 'optimum response time'. This procedure has precedence within the Irish context as the SEC (2018) have included 'suggested maximum response times' in the paper-based Junior Certificate Maths exam. It is unsurprising that Piech and Gregg (2018) are interested in investigating time management strategies for online CS exams. Consultations with CS instructors in Dublin City University have indicated that time management is often a cause for concern when students are writing code (handwritten or typed) in timed examinations (Source: Day, Personal Communication). The iterative nature of writing code can often be 'all consuming'. Students find it difficult to 'forget about it and move on' which must be done in an examination in order to maximise marks and achievement. A timer like this could be a valuable support to students in a CS exam.

6.4 Guidelines for CS Examination: Test Design

- Guidelines and strategies to assist in the construction of MCQs that can measure higher and lower order thinking skills e.g. verb selection etc., should be consulted.
- Text, images and animations can all be used as question prompts for MCQs and can provide a broader representation of the targeted skills, knowledge and abilities being assessed.
- Figural response items (e.g. drag-and-drop) are considered 'as good as' traditional text-based MCQs and would be a valuable inclusion to the upcoming CS exam. They would be particularly well suited to certain CS tasks (e.g. Parson's Puzzles).
- Non-digital resources to support novice programmers should also be considered e.g. a 'log table' for CS (similar to those used by students in Maths exams).
- If students are required to write code in the 2020 CS exam, then, based on the experiences of other educational institutions, a modified text editor that supports **auto-indentation, line-numbering and syntax highlighting**, which students should be familiar with from classroom activities, would be best suited to this task. If an IDE that includes the ability to compile and execute code is used for the 2020 CS exam, procedures need to be in place to support students' efficient use of these features e.g. guidelines on how many lines of code their answer should contain.
- Anecdotal evidence suggests that timing in CS exams can be an issue. As a result, the order of and time allocation for questions in the CS exam should be carefully considered (e.g. all code writing tasks in a separate section or paper). An 'optimum response time' could be allocated to particular sections (e.g. MCQ) or coding questions to ensure that students complete the exam in enough time. It would be advisable to explore these issues in advance of the 2020 exam day in a field trial to determine what approach would be most beneficial to students.

7.0 Test Deployment and Delivery

Schools will be the most likely setting for the CS CBE. Concerns surrounding the school-based delivery of the high-stakes CS exam in Ireland should be addressed by consulting literature from other countries or institutions who have previously trialled the implementation of CBEs within second-level settings. Three specific case studies from

New Zealand, the United States and the Programme for International Student Assessment (PISA) have been identified as suitable sources of information for the Irish CS CBE. Given their relative similarity to the Irish education system and the range of literature available, New Zealand's experiences developing CBEs should be closely examined. Their original investigation into CBEs with the 2013/14 electronic Mathematics Common Assessment Task (eMCAT) should yield some useful information that could be applied to any possible field trial that may be conducted in an Irish setting. The country's current efforts to develop and deliver their National Certificates of Educational Achievement (NCEA) using a long term digital platform should also be studied to identify up-to-date practices on the use of CBEs in second-level schools. Studies regarding the rollout of the digital version of the Partnership for Assessment of Readiness for College and Careers (PARCC) from the past five years in the US should also be considered for further advice and guidance. Research by PISA on the deployment of the 2015 CBE that was used in over 72 countries with 14-15 year olds can be analysed for the same purpose.

7.1 New Zealand

7.1.1 eMCAT Project 2014

The process of creating an online platform to deliver high stakes examinations to second-level students has been ongoing in New Zealand for some time and began with their development of the 2013/14 electronic Mathematics Common Assessment Task (eMCAT). The Mathematics Common Assessment Task (MCAT) is a paper based assessment developed for the assessment of a particular achievement standard (algebraic problem solving) within New Zealand's second-level curriculum. It is a stand-alone assessment event and has relatively low stakes. The New Zealand Qualifications Authority (NZQA) conducted a pilot CBE using the MCAT in 2013 with 27 schools and 2,470 students (aged 14-16). 146 students completed a qualitative survey to give more detailed feedback to test-developers. The NZQA (2014) hoped to gain a better understanding of the procedures and best practice guidelines required to develop and deliver a CBE in New Zealand as well as the necessary familiarisation and socialisation activities that should be undertaken within the New Zealand education community to ensure the seamless transition to digital assessments. The 2014 eMCAT pilot resulted in some key lessons for the NZQA in a number of areas as outlined by Table 4 (overleaf).

Table 4 Key Findings from the eMCAT Report (NZQA, 2014)

Area	Key Findings
CBE Interface	- 'An examination set out for an A4 computer format does not suit a computer screen' (NZQA, 2014, p. 6) - Issues regarding navigation, scrolling, question presentation. - Negative feedback from students and teachers
Familiarisation Activities	- Students should be 'acclimatised to assessment in an electronic environment' to ensure that they were comfortable undertaking a CBE (NZQA, 2014, p. 10). - Candidates set up an account with the CBE provider and were then granted logon details so they could access the eMCAT familiarisation activities to allow them some time to 'get used to' the interface. - Students felt that they did not have adequate time or were not supported by schools to complete the available familiarisation activities.
Exam Security and Technical Issues	- Only minimal security procedures were in place. - A two-factor authentication process for future assessments to ensure validation of the candidate. - Features to block the use of certain online tools (e.g. Google) was suggested based on teacher/ invigilator feedback. - Lots of technical issues were noted in relation to internet capacity. Some school systems could not handle a direct web-based logon process by large numbers of students.
Marking	- A hybrid marking system was used where the computer gave a recommended result to the marking teacher. The teacher could accept or reject this mark - This feature was not positively received by teachers as the marking interface was not optimised for the correction process. Teachers could not build up a 'marking flow' as they could with paper-based assessments

In 2015, a secondary project was undertaken to build on the learning from this eMCAT trial involving the subjects of Maths, Science and French (NZQA, 2015). The user interface was heavily modified in accordance with the feedback received from students. Students undertaking any of these trial CBEs (*n*=17,106) were set up with an account well in advance of their examination dates to allow them to engage with the requisite familiarisation activities. The familiarisation activities were available to students undertaking a Maths CBE approximately 6 weeks before the exam. Familiarisation activities for the Science and French CBEs were available between 4 and 6 weeks before the set exam date (NZQA, 2015). Taking into consideration the recommendations from the eMCAT pilot, the following security additions were added to complement standard invigilation:

- Candidates were required to take the CBE in 'full-screen' mode.
- A two-step (username and password) authentication system was used.
- If students accessed a file or application that was not required, a report was sent to the examiner and a dialogue box appeared on the student's screen encouraging them to return to their assessment.

While a slightly higher rate of satisfaction with these CBEs was noted in comparison with the 2014 eMCAT, student and teacher satisfaction with the CBEs was still relatively low. Interestingly, student experiences of CBEs tend to differ depending on the subject. Student feedback from the pilot indicated that students had a better testing experience with CBEs for language-based subjects. In this study, 31% of students who completed the survey at the end of the eMCAT were asked to compare it to their experiences of PBEs. 30% judged it as neither better nor worse and 60% of respondents indicated they felt it was worse or much worse. While these subject based preferences could be attributed to the 'disconnect' found between the teaching approach used in classrooms for mathematics (which is very paper-based; NZQA, 2015) and the assessment approach, this discrepancy could, in part, be explained by the 'technical difficulties' experienced by many students when trying to record their workings on a computer screen or tablet. Students and teachers offered further suggestions for any future CBEs in New Zealand. The key recommendations from this follow-on study were as follows:

- Testing authorities should be aware of the profile of various devices and operating systems available in schools or among test-takers before developing any administration guidelines for schools.
- A clear set of technical specifications needs to be available to schools to ensure that any school-based devices meet the minimum technical requirements for an optimal testing experience and that adequate infrastructure is available.
- The UI of any online examination should allow students to note-take and work 'on' their answers before submission. If certain symbols need to be used, these should be easy to access and search for.
- Markers were much happier with the new user interface for marking scripts as it allowed 'strip marking' (where the same question from different students was marked sequentially).

7.1.2 NCEA Digital Assessment Trials and Pilots 2016-2018

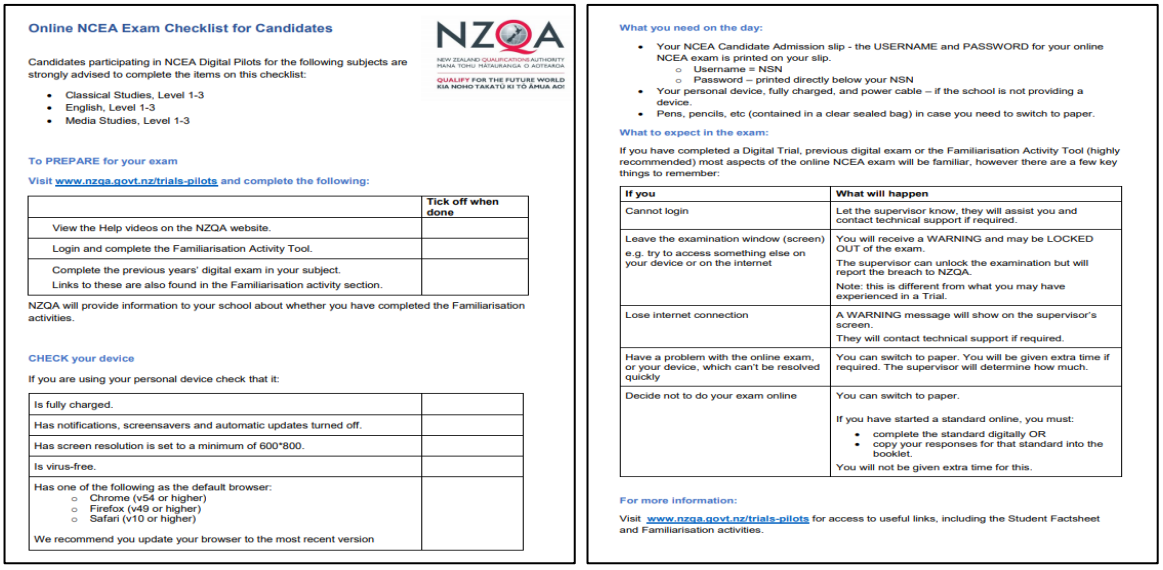
New Zealand's National Certificates of Educational Achievement (NCEA) are the main national qualifications for senior secondary school students. Each year, students study courses or subjects that are assessed against a number of standards. When a standard has been reached, the student gains credits which contribute to the NCEA certificate achieved. Similar to the Leaving Certificate in Ireland, the NCEA is used for selection by universities and is recognised by employers. Managed by NZQA (2018a), the country aims to have all NCEA examinations available on a long-term online platform in 2020. Taking into consideration the valuable lessons learned from the eMCAT project, the NZQA began the multi-year NCEA Digital Trials and Pilots scheme in 2016 where digital NCEA examinations in a range of subjects were developed, deployed and delivered to large groups of students. Schools had a choice of implementing a trial CBE (a practice assessment that does not contribute to NCEA credits) or a pilot CBE (which does contribute to their CBE credits). The aim of the Digital Trial and Pilot examinations is to help schools and students make the transition to full digital assessments by 2020. It enables schools to 'test their readiness, and provide an opportunity for students to experience assessment in a digital format' (NZQA, 2018a). In this project, each CBE is evaluated and reported on an annual basis to demonstrate how student and teacher feedback will contribute to the refinement of the next year's CBE. Over the last three

years, the NZQA have gained significant experience regarding the practicalities associated with the deployment and delivery of high-stakes CBEs in second-level settings. Key recommendations in relation to familiarisation activities, school-based preparation, exam day procedures, security and the value of user feedback can be extracted from this research to inform Ireland's deployment of a CBE for CS.

7.1.2.1 Key Recommendations: New Zealand

7.1.2.1.1 Familiarisation Activities

- Familiarisation Activities allow students to navigate through the structure of a digital examination (e.g. logging in, submitting answers etc.,). They should include tutorials and videos on how to use the features found in the digital examinations (e.g. word processing tools etc.,). The NZQA (2017) ensured that these familiarisation tools are available to candidates **at least** 6 weeks before exam day.
- Students should engage with these familiarisation **activities at least once** prior to examination day. Feedback from the 2017 student focus group indicated that familiarisation activities were helpful in preparing them for the official CBE.
- A **candidate checklist** such as the one provided on the NZQA website (2018b; Figure 10; Appendix 1) should be available to students. It will remind students to complete any familiarisation activities, check if their device meets the minimum technical requirements and is prepared for examination use. A list of 'what if' scenarios will ensure that candidates are aware of pertinent procedures.



Online NCEA Exam Checklist for Candidates

NZQA
NEW ZEALAND QUALIFICATIONS AUTHORITY
MANA TOHU MATAURANGA O AOTEAROA
QUALIFY FOR THE FUTURE WORLD
KIA NOHO TAKATŪ KI TŌ ĀMUA AO!

Candidates participating in NCEA Digital Pilots for the following subjects are strongly advised to complete the items on this checklist:

- Classical Studies, Level 1-3
- English, Level 1-3
- Media Studies, Level 1-3

To PREPARE for your exam

Visit www.nzqa.govt.nz/trials-pilots and complete the following:

	Tick off when done
View the Help videos on the NZQA website.	
Login and complete the Familiarisation Activity Tool.	
Complete the previous years' digital exam in your subject. Links to these are also found in the Familiarisation activity section.	

NZQA will provide information to your school about whether you have completed the Familiarisation activities.

CHECK your device

If you are using your personal device check that it:

Is fully charged.	
Has notifications, screensavers and automatic updates turned off.	
Has screen resolution is set to a minimum of 600*800.	
Is virus-free.	
Has one of the following as the default browser: ○ Chrome (v54 or higher) ○ Firefox (v49 or higher) ○ Safari (v10 or higher) We recommend you update your browser to the most recent version	

What you need on the day:

- Your NCEA Candidate Admission slip - the USERNAME and PASSWORD for your online NCEA exam is printed on your slip.
  - Username = NSN
  - Password – printed directly below your NSN
- Your personal device, fully charged, and power cable – if the school is not providing a device.
- Pens, pencils, etc (contained in a clear sealed bag) in case you need to switch to paper.

What to expect in the exam:

If you have completed a Digital Trial, previous digital exam or the Familiarisation Activity Tool (highly recommended) most aspects of the online NCEA exam will be familiar, however there are a few key things to remember:

If you	What will happen
Cannot login	Let the supervisor know, they will assist you and contact technical support if required.
Leave the examination window (screen) e.g. try to access something else on your device or on the internet	You will receive a WARNING and may be LOCKED OUT of the exam. The supervisor can unlock the examination but will report the breach to NZQA. Note: this is different from what you may have experienced in a Trial.
Lose internet connection	A WARNING message will show on the supervisor's screen. They will contact technical support if required.
Have a problem with the online exam, or your device, which can't be resolved quickly	You can switch to paper. You will be given extra time if required. The supervisor will determine how much.
Decide not to do your exam online	You can switch to paper. If you have started a standard online, you must: • complete the standard digitally OR • copy your responses for that standard into the booklet. You will not be given extra time for this.

For more information:

Visit www.nzqa.govt.nz/trials-pilots for access to useful links, including the Student Factsheet and Familiarisation activities.

Figure 10 Online NCEA Exam Checklist (NZQA, 2018b) (Appendix 1)

7.1.2.1.2 School-Based Preparation

- Schools should be aware of the technical and digital examination supports required to allow their students to participate in digital examinations. Issues relating to the following should be considered before delivering a CBE according to the NZQA (2018c):
    - Network and internet connection
    - Device specifications
    - Browser options
    - Device preparation
    - Familiarisation activities
    - Examination room set-up
    - Examination supervision
    - Marking
- Schools should be given an opportunity to 'walk through their own specific scenarios' with the testing authorities in advance of the examination (NZQA, 2017, p. 16).
- Schools should consider a number of factors beyond a student's device specifications for a CBE. For example, access to appropriate work spaces for students under examination conditions (e.g. with reduced opportunities for screen peeking, managing power supplies, managing temperature) should also be considered (NZQA, 2017) when selecting an examination room.
- Teachers, invigilators and examiners should all engage in some form of training in advance of the examinations to ensure that they are aware of the procedures and unique challenged involved with CBEs.

7.1.2.1.3 Examination Day

- A helpdesk should be available to each school so that there is a point of contact to assist in any technical difficulties (e.g. students being erroneously locked out of testing software).
- Contingency plans should be in place if there is a digital failure (e.g. a paper-based exam, a start-time 'window' rather than an official start time, spare devices etc.,).

- NZQA provided schools with student logins and passwords for the Digital Trials in an Excel spreadsheet. Schools should decide if these are given to students in before the examination day or on the morning of the examination. Exam Centre Managers in New Zealand found that the distribution of passwords etc., on the morning of the exam was quite cumbersome (NZQA, 2018d). If they are given to students prior to examination day, then procedures should be in place if students lose their login details.

7.1.2.1.4 Security

- Access to a virtual dashboard to monitor student activity (e.g. student progress, view of a student's screen) is extremely valuable to examiners and invigilators and should be a key component of the invigilation process (NZQA, 2017).
- The testing platform provided students with their two-step log-in details (username, password) which was distributed within schools.
- Access to an unauthorised application or programme resulted in a student warning via dialogue box and the invigilator being alerted. If the student did not return to the examination they were 'locked out' and required support from the invigilator.

7.1.2.1.5 User Feedback

- User feedback is vital to the successful development of CBEs. The NZQA have developed a series of surveys and questions for students, teachers, invigilators, principals and exam markers to gain insight into the experiences of these stakeholders using a CBE in a high-stakes setting (NZQA, 2018a). This feedback has been used to inform subsequent iterations of CBEs in New Zealand. Focus groups and interviews have also been introduced to the feedback process associated with the NCEA CBEs (NZQA, 2017; 2018b).

According to the NZQA (2018a), almost three quarters of New Zealand secondary schools and around 30,000 students have experienced at least one online examination since 2014. In their latest review of the use of CBEs in schools, the NZQA (2018e; n=990)

found that more students strongly agreed that CBEs gave them a positive examination experience (54%) than agreed (44%). Levels of satisfaction with this mode of assessment have been steadily rising since the 2014 eMCAT project. However, there have been some recent issues with these CBEs that have received some media attention in New Zealand. For example, a server hosting an NCEA English CBE 'crashed' two hours into a three-hour exam, causing more than 3600 students significant distress (SchoolLeaver, November 2018). In 2017, approximately 358 students were mistakenly given fail grades by the online marking system, an error that was not noticed until months later (Radio New Zealand, February 2018). These 'technical difficulties' highlight the need for constant vigilance in relation to the overall maintenance of the digital infrastructure, even when the initial development has been completed.

7.2 United States (US)

The Partnership for Assessment of Readiness for College and Careers (PARCC; 2018a) is a group of states working together to develop assessments that measure whether students in the US from Grades 3-11 are 'on track' to be successful in college and careers. These high-stakes exams, which are used to evaluate teachers and schools in some states, measure students' knowledge of grade-level content in English language arts and mathematics as defined by state standards. In spring 2014, more than one million students in 16,000 schools participated in the field testing of new CBEs developed by PARCC. The field tests were conducted 'to ensure the assessments are valid and reliable, while providing an opportunity for state and local administrators to gain valuable insight into the effective management and use of new technologies to support computer-based testing' (Rennie Center, 2015, p. 1) before the formal spring 2015 CBEs.

While a complete review of this field trial was completed and published by PARCC (2014), some states engaged in an independent evaluation of this field trial, including Massachusetts (Rennie Center, 2015). A case study of two school districts' (Burlington [6 schools; 2,200 students], Revere [3 schools; 950 students]) experiences in implementing this field trial were recorded. This case study, along with PARCC's (2014) review can be used to identify what challenges students and schools involved with this field trial experienced in order to determine best practice guidelines for the deployment and delivery of CBEs in relation to Ireland's CS exam. In particular, the challenges associated

with Technical Infrastructure as well as Student Preparedness and Experience will be presented along with recommendations to address these issues so that they do not occur in other field trials.

7.2.1 Technical Infrastructure

In the Massachusetts case study, the two school districts had both invested a large amount of time, effort and finances to upgrade the technical infrastructure of their schools (Rennie Centre, 2015). As a result, both districts had fibre optic connections between school buildings that provided high-speed internet service. This connection ensured that the schools had sufficient bandwidth to easily access downloaded test materials at each school. Schools in both districts provided the devices student would use in the test. However, schools in Burlington and Revere found that test materials tended to freeze or spool for extended periods of time when the downloaded materials were initialised for use on examination day. Although internet connectivity and inadequate local technology were initially blamed for such problems, the Rennie Centre did not consider this a viable reason given the relative ease that items were downloaded in both districts, and the extensive preparation conducted by staff in schools. In advance of the PARCC field test, the Burlington and Revere districts engaged in excellent preparatory activities (Table 5; Rennie Centre, 2015, p. 6):

Table 5 Overview of school-based preparatory activities

Activity	**Actions**
Equipment Inventory	District-level IT staff inventoried devices and determined if they met PARCC specifications (screen size, processing speed etc.,).
Device Configuration	Devices were upgraded to latest browser versions by school staff. Relevant software (anti-virus etc.,) was also updated.
Device Preparation	IT staff installed CBE-specific software (TestNav8) for test administration on all suitable devices. PARCC made this software available to download through the CBE provider.
Connectivity Test	Once all updates were completed, IT staff tested all devices to determine if internet service was adequate for testing.

These extensive surveys of technical infrastructure were relatively common in schools. Based on survey feedback collected by PARCC (2014), 69% of test administrators and test coordinators conducted some form of infrastructure trial. Despite this, challenges at local level still existed. The freezing and excessive spooling of examination materials were attributed to the interaction of the test materials with the testing platform. The Rennie Centre (2015, p. 4) noted that many resources and activities related to the testing platform were released late which raised 'questions about whether products had been sufficiently tested for the scale at which they were being used' or on the range of devices available in schools. As this experience was replicated in many schools and districts (PARCC, 2014), the PARCC consortium has provided additional supports and to improve the local technology experience. These guidelines are used to inform current PARCC CBEs.

- PARCC release technology guidelines to help schools, districts, and states determine the level of readiness of their existing devices inventories and the new hardware they may need to purchase. These manuals allow schools to identify what devices meet PARCC's minimum requirements for CBEs (e.g. PARCC, 2018b). They also provide clear checklists for schools so that teachers are aware of the steps required to prepare a school for a CBE.
- PAARC organise 'Infrastructure Trials' in advance of examination day. Schools can conduct a test-day network use simulation so schools can determine if they have sufficient bandwidth. Students can complete familiarisation activities during this time and teachers and invigilators can practice invigilation procedures if a device fails or if something refuses to load on a device. When CBEs were first introduced to PARCC, these trials were available 3-4 months prior to testing. Now they are available to use throughout the year at the school's discretion.
- PARCC delivers training modules and has several tutorial videos available to guide users in conducting an infrastructure trial. These modules aim to help schools understand, manage, and make decisions in preparing school technology to be used for online testing (PARCC, 2014).

7.2.2 Student Preparedness and Experience

The 2014 PARCC field test enabled students and teachers in Burlington and Revere to make preliminary observations about student preparation for a computer-based PARCC tests. According to the Rennie Centre (2015), educators reported investing some class time to guide their students through the use of the publicly-available PARCC preparation activities and materials. According to survey data, approximately 75% of the teachers surveyed indicated that they used class time to have students watch the PARCC tutorial with around 70% reviewing the publicly-released test items with their students. As was seen in New Zealand, familiarisation activities are essential to ensure that students have enough time to become familiar with this examination environment. Based on user feedback, familiarisation activities for the 2015 spring PARCC exam were available 3-5 months in advance of this testing period. A range of familiarisation activities for current PARCC exams are available online throughout the year (PARCC, 2018c).

In the case of Massachusetts, student opinions from focus groups indicated that the practice items that they did in class were 'not very helpful as the content of these did not seem representative of questions on the actual test' (Rennie Centre, 2015, p. 14). The experiences of these students aligned well with the national experience for students who undertook the Mathematics PARCC assessment in the field trial (PARCC, 2014). Access to and use of particular tools (e.g. calculator, highlight tool, text-to-speech etc.,) rather than the actual input of their answers seemed to have cause some difficulty amongst students. Based on the feedback received from students and teachers in the field trial, PAARC (2014) and school districts in Massachusetts recommended that an increase in the use of technology in classrooms would ensure that their students are comfortable learning and demonstrating their work in digital environments (Rennie Centre, 2015). Furthermore, ensuring that student use of technology is incorporated into every lesson would enhance student familiarity with CBEs. Teachers and PARCC administrators/ invigilators also engaged in the following preparatory activities to ensure that they were familiar with the procedures involved in CBEs:

- Regional Training workshops
- Site Visits from PARCC
- Question and Answer Workshops

- Online training modules

As demonstrated here, field trials provided the PARCC testing authorities with vital information that gave some guidance as to the best ways to deploy the test for the formal exam and to prepare the relevant stakeholders for this change. The insights that emerged from this field trial differ to those that came from the New Zealand field trials, although some similarities are present. This highlights the importance of interpreting and applying a country's experiences with CBEs with some caution as they are contextually bound.

7.3 PISA 2015

The Programme for International Student Assessment (PISA) is an international assessment of 15-year-old students that measures how well students apply their knowledge and skills to solve problems related to reading, mathematics, and science in real-life contexts. This project by the OECD involves over 70 countries and is the largest of its kind. In 2015, approximately 60 countries, including Ireland, transitioned entirely to computer based assessment. Prior to this, each of these countries carried out a field trial in March 2014, **one year in advance** of the main study. Approximately 2,000 students in 25 schools were involved in this field trial in Ireland, representing approximately 40% of the sample that would be used in the actual study in March 2015 (Educational Research Centre [ERC], 2014). 80 students in each school participated in the study, where they were assigned to one of three groups, two of which used CBEs. The goals of the 2014 Field Trial were to identify optimal operational procedures for school-based use of CBEs and to examine the quality of the newly developed items for computer-based delivery. As this is an international assessment the content, organisation and deployment of PISA tests are significantly different from those used in the Leaving Certificate. However, as this assessment did involve the use of a CBE field trial in Ireland within the past 5 years, it could offer some information to the relevant authorities who wish to deploy a field trial in advance of the 2020 CS exam.

Shiel, Kelleher, McKeown and Denner (2015) referred to this field trial in their analysis of Ireland's PISA 2015 results. According to Shiel et al. (2015, p. 61), the computer-based assessment 'found a decline in item omissions... and the position effect

was reduced by between one-third and one-half on CBA [computer-based assessment] compared with PBA [paper-based assessment]'. The CBE also provided better data quality as timing data allowed for a clearer distinction to be made between omitted and not reached items. While this gave some reassurance to the reliability of CBEs, the field trial also gave the Education Research Centre (ERC; the institution which implements PISA on behalf of the Department of Education and Skills) clear guidelines for the implementation of a CBE in Irish schools for the actual study in March 2015.

For example, one of the recommendations from the PISA 2015 field trial was to administer the assessment on laptops provided by the ERC to schools for the assessment. Site visits conducted in schools for the field trials found that the majority of school devices did not meet the minimum PISA specifications. As a result, schools or students were not responsible for providing their own devices for the actual March 2015 test. Instead, a laptop hire company provided 800 laptops to the ERC, onto which the assessment was loaded. The ERC was responsible for the transport of this equipment to the school. A technician accompanied these devices and these individuals would attend to any 'technical difficulties' that arose e.g. log-on issues, drive failure etc., (Shiel et al., 2015). The CBE was administered from each laptop's hard drive, rather than from a USB stick (as had been the practice in the field trial). Shiel et al. (2015, p. 22) noted that this 'seemed to improve the speed at which students' accessed to the material'. However, it is important to note that while the field trial did take place in schools, it was entirely managed and organised by the ERC. Furthermore, the administration and security requirements for PISA are likely to be different to those in the Leaving Certificate so it may not be possible to 'map' these procedures onto any CBE trials for the Leaving Certificate.

Contrary to the procedures in New Zealand and the US, no familiarisation activities for the PISA 2015 CBE were available to students in advance of the actual study in 2015 in the field trial the previous year. Instead, a general orientation screen is introduced to students at the start of the actual CBE test. This orientation introduces students to 'the screen design and those response modes that were common' (OECD, 2017, p. 46). Given the low-stakes nature of the test for students (although PISA has much higher stakes for countries at a policy level), the OECD considers this sufficient for students to gain an

experience of the interface. High stakes exams would likely require more familiarisation activities to ensure that students are comfortable in this testing environment.

The procedures employed by the ERC on behalf of the OECD for the main study of PISA 2015 are unlikely to be used in any CBEs for the Leaving Certificate. However, they do show the value of running a field trial well in advance of any official exam days. Doing so ensures that the relevant testing authorities can be confident of any operational procedures that they recommend.

7.4 Guidelines for CS Examination: Test Deployment and Delivery

- A field trial should be conducted in advance of the CS exam in June 2020 to test the user interface and to determine the key technical, logistical and security issues around school-based CBE delivery in Ireland. This field trial should be conducted within a timeframe that allows test developers to modify the test design and administration procedures according to user feedback while also ensuring that students have access to familiarisation activities based on the final UI 1-5 months in advance of exam day.
- User feedback from these field trials will be essential. This can take the form of observations, surveys and focus groups involving relevant stakeholders.
- Technology guidelines should be distributed to students and schools to help them to make technology decisions to best meet the instructional and assessment needs of the CS exam. These guidelines should contain the minimum device and infrastructural requirements to deploy the CS exam in 2020. These should be distributed well in advance of the exam date to ensure that students and schools have adequate preparation time.
- Schools will need extra support to ensure that they are prepared to deliver a CBE. Site visits from technical advisors to participating schools should be provided well in advance and just prior to the exam date. This will allow schools an opportunity to relay any local, school-based issues that may impact on students' testing experiences. These advisors should give guidelines around room set up, contingency plans etc., A helpline that schools can contact directly on the examination day should also be available.

- Schools should make contingency plans in the event of full or partial technology failure. These should be agreed with the relevant testing authorities and students should also be made aware of them.
- The orientation of students to a novel mode of assessment, particularly when it represents a major departure from previously established norms, is a key issue. Instructors should integrate the assessment method at an early point in the curriculum in order to reduce student anxiety. Therefore, familiarisation activities for all question types that students may encounter in the CS exam should be available to students. Reports from New Zealand and the US indicate that such activities are available at least **1-5 months before exam day**.
- Professional development for teachers and exam invigilators should be provided. This can take the form of face-to-face workshops or online tutorials.
- Two-step verification sign-ins are used in other high-stakes exams in second level settings. This is the minimum requirement that should be used in the 2020 exam.

8.0 Key Questions

The provision of a CBE to assess student attainment in the newly developed Leaving Certificate subject of Computer Science by June 2020 is a significant undertaking for all involved. To support key stakeholders in their efforts to prepare for this task, the current report aimed to provide an overview of current issues pertaining to the use of CBEs in second-level schools using relevant literature from international research, testing organisations and other education systems. Discussions surrounding test mode comparability, human-computer interaction, test design and the organisation of field trials have all been presented. Therefore, this report should offer some guidance to all relevant stakeholders on how an effective CBE can be designed and deployed in Ireland. It should be used to inform any decisions these key stakeholders make regarding their planned implementation of this CBE.

In particular, it is hoped that the current report will inform the NCCA's attempts to contribute to the debate around the design and delivery of this CBE. A number of questions must first be considered before beginning the process of developing a CBE for CS. Table 6 (overleaf) organises these questions into the themes highlighted in the report.

Table 6 Key Questions for the 2020 CS exam

Area	Key Questions	Factors to Consider
Test Mode Comparability	- Will both a paper-based and computer-based exam be available? - If both versions are available, will students get to choose which version they use? - Will the entire CS exam be presented in one format? - Will students, schools or the SEC provide the devices for CBEs? - What are the minimum specifications (e.g. screen size, keyboard, processing power etc.,) required for the devices needed for the planned CBE?	- If two versions of the exam are made available, a comparability study is recommended. - If the test is to be device agnostic, clear guidelines regarding device specifications should be provided to students and schools as soon as possible. - If a range of devices are permitted to be used for the CBE, care must be taken to ensure that there is a consistent user experience between devices for students.
Human-Computer Interaction	- What should the User Interface of the planned CBE look like? - How much freedom will students have to navigate within the exam before, during and after they submit their answers? - What tools should be included in the CBE? Annotation tools? Word processing tools? A timer? - Will exam markers receive training about the unique scoring challenges associated with a CBE?	- The User Interface should be designed in accordance with Nielson and Molich's (1990) work. - Freedom to navigate within the exam will be dependent on the overall layout of the exam. - If word processing tools are to be include in the interface, then they should resemble software that students are already familiar with. - Specific training should be provided to exam markers for a CBE.

Test Design	- What type of stimuli (e.g. video, images etc.,) could be used as prompts for MCQs? - How can MCQs target lower and high-order skills? - What other alternatives are there to MCQs? How should these be designed? - What features should be included in a sandbox? - Should a text-editor or IDE be used in the CS CBE? - How can concerns about student time management skills when writing code be addressed?	- Different stimuli in MCQs elicit different skills. - Guidelines to write MCQs should be consulted. - Some aspects of CS may be particularly well suited to figural response items (e.g. Parson's Puzzles). - IDEs in timed examinations may not be appropriate. - Text-editors should include certain features (e.g. indentation, colour coding) to support novice programmers in line with UI principles. - An off-line resource akin to a log book or 'cheat sheet' could also be used to support students in exams.
Delivery and Deployment	- When should the first Field Trial be conducted? - How large should the Field Trial be? - How will user feedback be gathered? - When will students have an opportunity to become familiar with the CBE interface? - What type of supports will be made available to schools (before and during) the June 2020 exam to ensure that technical infrastructure is appropriate? - What form of training and support will teachers, exam invigilators and exam markers receive? - What security features will be included in this CBE?	- Students should engage in some familiarisation activities 1-5 months before examination day. This should inform the dates of the planned field trials. - User feedback can be gathered using surveys and interviews. Cognitive labs may also be used. - Technical infrastructure in schools should be surveyed well in advance of and just prior to the exam. Schools should be able to conduct an 'infrastructure trial' before exam day. Each school should be able to access a technician or helpline when needed.

Reference List

Apple. (2018). *UI design: Do's and don'ts*. Retrieved April 18th 2018 from https://developer. apple.com/design/tips/

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruickshank, K.A., Mayer, R.E., Pintrich, P.R., … Wittrock, M.C. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of Educational Objectives.* (Complete edition). New York: Longman.

Arkes, H.R. & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behaviour and Human Decision Process, 35*(1), 124-140. https://doi.org/10.1016/0749-5978(85)90049-4

Backes, B. & Cowan, J. (2018). *Is the pen mightier than the keyboard? The effect of online testing on measured student achievement* (Working Paper 190). National Centre of Analysis of Longitudinal Data in Educational Research. Retrieved 6th November from https://caldercenter.org/sites/default/files/WP%20190.pdf

Barros, J.P. (2018). Students' Perceptions of Paper-Based vs. Computer-Based Testing in an Introductory Programming Course. In *Proceedings of the 10th International Conference on Computer Supported Education (CSEDU 2018).* Retrieved December 2nd, 2018 from http://www.scitepress.org/Papers/2018/67942/67942.pdf

Bennedsen, J., Caspersen, M., & Allé, F. (2007). Assessing process and product. *Innovation in Teaching and Learning in Information and Computer Science, 6*(4), 183-202. https://doi.org/10.11120/ital.2007.06040183

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PLoS ONE*, *10*(12), 1–13.

Bridgeman, B., Lennon, M.L. & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16*, 191–205. doi:10.1207/S15324818AME1603_2

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research & Evaluation*, *22*(1). Retrieved from https://pareonline.net/getvn.asp?v=22&n=1

Cantillon, P., Irish, B., & Sales, D. (2004). Using computers for assessment in medicine. *British Medical Journal (Clinical Research Edition)*, *329*(7466), 606–609. https://doi.org/10.1136/bmj.329.7466.606

Chen, L.J., Ho, R.G., & Yen, Y.C. (2010). Marking strategies in metacognition-evaluated computer-based testing. *Educational Technology & Society*, *13* (1), 246–259. Retrieved October 25th 2018, from

https://pdfs.semanticscholar.org/2267/05f01d441f079307e3b313c7a5d73ecb6838.pdf

Conn, A.P. (1995). Time Affordances: The Time Factor in Diagnostic Usability Heuristics. In *Proceedings of the 1995 SIGCHI Conference on Human Factors in Computing Systems*. (pp. 186-193). ACM Press: New York.

Conrad, F., Couper, M., Tourangeau, R. & Peytchev, A. (2010). The impact of progress indicators on task completion. *Interacting with Computers, 22(5)*, 417-427.

Csapó, B., Ainley, J., Bennett, R., Latour, T., & Law, N. (2012). Technological issues of computer-based assessment of 21st century skills. In B.McGaw, P. Griffin, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). New York, NY: Springer.

Dadey, N., Lyons, S. & DePascale, C. (2018). The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice. *Applied Measurement in Education, 31(1)*, 30-50. doi: 10.1080/08957347.2017.1391262

Davis, L., Janiszewska, I., Schwartz, R. & Holland, L. (2016). *NAPLAN Device effects study*. Melbourne, Australia: Pearson. Retrieved 3rd March, 2018 from http://nap.edu.au/docs/default-source/default-document-library/naplan-online-device-effect-study.pdf

Davis, L., Kong, X. & McBride, Y. (2015). *Device comparability of tablets and computers for assessments purposes*. Retrieved March 5th, 2018 from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/efficacy-and-research/schools/Device%20Comparability-%20Score%20Range%20and%20Subgroup%20Analyses.pdf

Davis, L., Kong, X., McBride, Y. & Morrison K. (2017). Device comparability of tablets and computers for assessment purposes. *Applied Measurement in Education, 30(1)*, 16-26. https://doi.org/10.1080/08957347.2016.1243538

Davis, L. L., Morrison, K., McBride, Y. & Kong, X. (2017). Disaggregated effects of device on score comparability. *Educational Measurement: Issues and Practice, 36(3)*, 35–45. https://doi.org/10.1111/emip.12158

Davis, L., Orr, A., Kong, X. & Lin, C. (2015). Assessing student writing on tablets. *Educational Assessment, 20(3)*, 180-198. https://doi.org/10.1080/ 10627197.2015.1061426

Davis, L.L. & Strain-Seymour, E. (2013). *Keyboard interactions for tablet assessments*. Washington, DC: Pearson Education. Retrieved Nov 15th, 2017, from http://researchnetwork.pearson.com/ wpcontent/uploads/Keyboard.pdf

Delen, E. (2015). Enhancing a Computer-Based Testing Environment with Optimum Item Response Time. *Eurasia Journal of Mathematics, Science and Technology Education, 11*(6), 1457-1472. https://doi.org/10.12973/eurasia.2015.1404a

Dembitzer, L., Zelikovitz, S. & Kettler, R.J. (2017). Designing computer-based assessments: Multidisciplinary findings and student perspectives. *International Journal of Educational Technology, 4(3)*, 20-31.

Denny, P., Luxton-Reilly, A. & Simon, B. (2008). Evaluating a new exam question: Parsons problems. In *Proceedings of the 4th international workshop on computing education research*, (pp. 113-124). ACM: Australia. Retrieved October 4th, 2018 from http://delivery.acm.org/10.1145/1410000/1404532/p113-denny.pdf?ip=136.206.193.128&id=1404532&acc=ACTIVE%20SERVICE&key=846C3111CE4A4710%2E821500BF45340188%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1544434702_8ca73ee0639dbe6ff1d4db2500800578

DePascale, C., Dadey, N. & Lyons, S. (2016). Score comparability across computerized assessment delivery devices. *Council of Chief State School Officers: Texas.* Retrieved March 3rd 2018 from https://www.nciea.org/sites/default/files/pubs-tmp/CCSSO%20TILSA%20Score%20Comparability%20Across%20Devices.pdf

Department of Education and Skills (DES). (2015). *Digital Strategy for Schools 2015-2020: Enhancing teaching, learning and assessment.* Dublin: Stationary Office.

Dickinson, M. (2011, December 5th). *Writing multiple choice questions for higher-level thinking.* Learning Solutions Magazine. Retrieved October 30th, 2018 from http://www.learningsolutionsmag.com/articles/804/writing-multiple-choice-questions-for-higher-levelthinking

Dillon, E., Anderson, M., & Brown, M. (2012). Comparing feature assistance between programming environments and their effect on novice programmers. *Journal of Computing Sciences in Colleges, 27*(5), 69–77.

Eberhart, T. (2015). *A comparison of multiple-choice and technology-enhanced item types administered on computer versus iPad* (Doctoral dissertation), University of Kansas, Lawrence, KS. Retrieved March 11th, 2018 from https://kuscholarworks.ku.edu/bitstream/handle/1808/21674/Eberhart_ku_0099D_14325_DATA_1.pdf?sequence=1

Education Central (2018). *Students wrongly failed by digital NCEA exam after receiving no marks* [web article]. Retrieved 6th November, 2018 from https://educationcentral.co.nz/students-wrongly-failed-by-digital-ncea-exams-after-receiving-no-marks/

Educational Research Centre (ERC). (2014). *Information for schools: PISA Field Trial* (March 2014). Retrieved December 1st, 2018 from http://www.erc.ie/documents/p15brochuresch.pdf

Fulcher, D. (2003). Interface design in computer-based language testing. *Language Testing, 20(4)*, 384-408. doi: 10.1191/0265532203lt265oa.

Frein, S. T. (2011). Comparing in-class and out-of-class computer-based tests to traditional paper-and-pencil tests in introductory psychology courses. T*eaching of Psychology, 38*(4), 282–287.

Gipps, C.V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Routledge.

Graham, S., Harris, K., & Hebert, M. (2011). *Informing writing: The benefits of formative assessment. A Carnegie Corporation Time to Act report.* Washington: Alliance for Excellent Education. Retrieved October 8th, 2018 from https://www.carnegie.org/media/filer_public/37/b8/37b87202-7138-4ff9-90c0-cd6c6f2335bf/ccny_report_2011_informing.pdf

Gross, P. & Powers, K. (2005). Evaluating assessments of novice programming environments. In *Proceedings of the first international workshop on computing education research* (pp. 99–110). ACM: Australia. Retrieved 4th December, 2018 http://delivery.acm.org/10.1145/1090000/1089796/p99-gross.pdf?ip=136.206.237.150&id=1089796&acc=ACTIVE%20SERVICE&key=846C3111CE4A4710%2E821500BF45340188%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1543927453_8171286a8bd3b3efc161541ff9533b79

Gursoy, D. (2016). *Assessing novice programmers' performance in programming exams via computer-based test* (unpublished master's thesis). Netherlands: University of Twente. Retrieved November 23rd, 2018 from https://essay.utwente.nl/70114/2/GURSOY_MA_BMS.pdf

Hadadi, A., Luecht, R.M., Swanson, D.B., & Case, S.M. (1998, April). *Study 1: Effects of modular subtest structure and item review on examinee performance, perceptions and pacing.* Paper presented at the National Council on Measurement in Education Annual Meeting, San Diego, CA.

Haghighi, P.D., Sheard, J., Looi, C.K., Jonassen, D., & Ikeda, M. (2005). Summative computer programming assessment using both paper and computer. In *Proceedings of the 2005 conference on Towards Sustainable and Scalable Educational Innovations Informed by the Learning Sciences: Sharing Good Practices of Research, Experimentation and Innovation*, (pp. 67–75). Amsterdam: IOS Press.

Haigh, M. (2010). *Why use computer-based assessment in education? A literature review*. Retrieved October 4th, 2018 from http://www.cambridgeassessment.org.uk/research-matters/

Halpin, H. (January 4th, 2018). *'A digital revolution': Forty schools to offer Computer Science as Leaving Cert subject in September*. Retrieved from http://www.thejournal.ie/computer-science-leaving-cert-3781048-Jan2018/

Hardcastle, J., Herrmann-Abell, C.F., % DeBoer, G.E. (2017). *Comparing student performance on paper-and-pencil and computer-based tests*. Retrieved November 1st, 2018 from https://mcmprodaaas.s3.amazonaws.com/s3fs-public/Comparing

Way, W.D., Davis, L., Keng, L. & Strain-Seymour, E. (2015). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology in testing: Improving educational and psychological measurement* (pp. 260-284). UK: Routledge.

Harms, M. & Adams, J. (2008). *Usability and design considerations for computer-based learning and assessment.* Paper presented at the March 2008 Meeting of the American Educational Research Associations (AERA). Retrieved October 30th, 2018 https://pdfs.semanticscholar.org/3661/cbe8b6d8fec7ad37648164d02f2a80d47960.pdf

Herold, B. (2016). PARCC scores lower for students who took exams on computers. *Education Week*, *35*(20). Retrieved November 8th 2018, from http://www.edweek.org/ew/articles/2016/02/03/parcc-scores-lower-on-computer.html

Illinois State Board of Education. (2015). *2014-2015 PARCC Online Paper and Pencil Analysis.* Retrieved December 1st, 2018 from https://www.isbe.net/Documents/2015-parcc-paper-online-analysis.pdf

Jackel, D. (2014). Item differential in computer based and paper based versions of a high stakes tertiary entrance test: Diagrams and the problem of annotation. In T. Dwyer, H. Purchase and A. Delaney. *Diagrammatic representation and inference* (pp. 71-77). London: Springer.

Jerrim, J. (2018). *A digital divide? Randomised evidence on the impact of computer-based assessment in PISA*. United Kingdom: Centre for Education Economics. Retrieved 3rd September, 2018 from http://www.cfee.org.uk/sites/default/files/CfEE%20Digital%20Divide_1.pdf

Jimoh, R., Kehinde Shittu, A., & Kawu, K.Y. (2012). Students' perception of computer based test (CBT) for examining undergraduate chemistry courses. *Journal of Emerging Trends in Computing and Information Sciences 3*(2). Retrieved from http://www.cisjournal.org/ journalofcomputing/archive/vol3no2/vol3no2_2.pdf

Kallia, M. (2018). *Assessment in Computer Science courses: A literature review*. Retrieved 5th September 2018 from https://royalsociety.org/~/media/policy/projects/computing-education/assessment-literature-review.pdf

Karim, N.A. & Shukur, Z. (2016). Proposed features of an online examination interface design and its optimal values. *Computers in Human Behaviour, 64*, 414-422. doi: http://dx.doi.org/10.1016/j.chb.2016.07.013.

Keane, N. & McInerney, C. (2017). *Report on the provision of Courses in Computer Science in Upper Second Level Education Internationally*. Retrieved September 3rd, 2018 from https://www.ncca.ie/media/ 2605/computer_science_report_sc.pdf

King, L., Kong, X.J. & Bleil, B. (2011, April). *Does size matter? A study on the use of netbooks in K-12 assessment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Retrieved November 15th, 2018 from https://images.pearsonassessments.com/images/PDF/AERA-Netbooks_K-12_Assessments.pdf.

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. https://doi.org/10.1080/08957340802558326.

Kong, X., Davis, L., McBride, Y. & Morrison, K. (2018). Response Time Differences Between Computers and Tablets. *Applied Measurement in Education, 31(1),* 17-29. doi: 10.1080/08957347.2017.1391261.

Lehane, P. (September, 2018). *Does it matter what device is used to administer technology-based assessments?* Presented at the European Association of Test Publishers (E-ATP), Athens, Greece.

Lim, M.L., Ayesh, A., Stacey, M. & Tan, L.P. (2017). The effects of task demand and external stimuli on learner's stress perception and job performance. In G.B. Teh and S.C. Choy (Eds.) *Empowering 21st Century Learners through holistic and enterprising learning* (pp. 88-101). doi: 10.1007/978-981-10-4241-6_10.

Lister, R., Adams, E., Fitzgerald, S., Fone, W., Hamer, J., Lindholm, M., McCartney, R., Mostrom, J., Sanders, K., Seppal, O., Simon, B. & Thomas, L. (2004). *A multi-national study of reading and tracing skills in novice programmers*. In Working Group Reports from ITiCSE on Innovation and Technology in Computer Science Education (ITiCSE-WGR'04). 119–150. Retrieved October 23rd, 2018 from https://www.cs.auckland.ac.nz/~j-hamer/ITiCSE04-wg.pdf

Lopez, M., Whalley, J., Robbins, P., & Lister, R. (2008). *Relationships between reading, tracing and writing skills in introductory programming*. Retrieved 25th October, 2018 from https://opus.lib.uts.edu.au/bitstream/10453/10806/1/2008001530.pdf

Lorié, S. (2015). *Reconceptualizing score comparability in the era of devices.* Presentation at the Association of Test Publishers conference, Palm Springs, CA. Summary retrieved March 23rd, 2018 from http://innovationsintesting.org/atp2015/program-webcast-ignite4-reconceptualizing-score.aspx

Lotteridge, S., Nicewander, W.A. & Mitzel, H.C. (2010). Summary of the online comparability studies for one state's end-of-course program. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 13-33).

Washington, DC: Council of Chief State School Officers. Retrieved 3rd March, 2018 from https://files.eric.ed.gov/fulltext/ED543067.pdf

Luecht, R.M., & Sireci, S.G. (2011). *A review of models for computer-based testing*. US: College Board Research Reports: Retrieved 20th October, 2018 from https://files.eric.ed.gov/fulltext/ED562580.pdf

Malik, A. (2018). *BlueBook: Secure, electronic computer science exams* [webpage]. Retrieved December 1st, 2018 from http://malikaliraza.com/projects/bluebook/index.html

Malone, S. & Brünken, R. (2013). Assessment of driving expertise using multiple choice questions including static vs. Animated presentation of driving scenarios. *Accident Analysis and Prevention*, *51*, 112–119. https://doi.org/10.1016/j.aap.2012.11.003

Marshall, C. (1997). Annotation: from paper books to the digital library. In *Proceedings of the Second ACM International Conference on Digital Libraries* (pp. 131–140). ACM: Pennsylvania. doi: 10.1145/263690.263806

Martinez, M. (1991). A comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement, 28*, 131–145. Retrieved from http://www.jstor.org/stable/1434795

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). New Jersey: Lawrence Erlbaum Associates.

Microsoft. (2018). *User interface principles*. Retrieved April 18th 2018 from https://docs.microsoft.com/en-us/windows/desktop/appuistart/-user-interface-principles

Molich, R. & Nielson, J. (1990). Improving a human-computer dialogue: What designers know about traditional interface design. *Communications of the ACM 33* (March). Retrieved October 25th 2018 from, https://pdfs.semanticscholar.org/8e67/d5075db82691aad39743d3414019ab4e38c0.pdf

National Council for Curriculum and Assessment (NCCA). (2018). *Computer Science: Curriculum specification*. Retrieved November 3rd, 2018 from https://www.ncca.ie/media/3369/computer-science-final-specification.pdf.

National Council for Curriculum and Assessment (NCCA). (2007). *Assessment in the primary school curriculum: Guidelines for schools*. Dublin: Stationary Office.

National Center for Education Statistics (2012). *The Nation's Report Card: Writing 2011 (NCES 2012–470).* Washington: Institute of Education Sciences, U.S. Department of Education.

New Zealand Qualifications Authority (NZQA). (2018a). *Digital assessment: NCEA digital trials and pilots* [webpage]. Accessed November 29, 2018 from https://www.nzqa.govt.nz/about-us/future-state/digital-assessment-trials-pilots/.

New Zealand Qualifications Authority (NZQA). (2018b). *Online NCEA exam checklist for candidates.* Retrieved 28th November, 2018 from https://www.nzqa.govt.nz/assets/About-us/Future-State/2018-trials-and-pilots/Candidate-Checklist-Pilots.pdf

New Zealand Qualifications Authority (NZQA). (2018c). *Technical requirements and other considerations* [webpage]. Retrieved 28th November, 2018 from https://www.nzqa.govt.nz/about-us/future-state/digital-assessment-trials-pilots/technical-requirements/

New Zealand Qualifications Authority (NZQA). (2018d). E*xam centre manager/supervisor survey: Digital pilots 2017.* Retrieved 29th November, 2018 from https://www.nzqa.govt.nz/assets/About-us/Future-State/2017-trials-and-pilots/ECMorSupervisorsurveyanalysisfinal.pdf.

New Zealand Qualifications Authority (NZQA). (2018e). *National Certificate of Educational Assessment (NCEA) external digital assessment 2017 trials and pilots: User experience evaluation report.* Retrieved November 28th, 2018 from https://www.nzqa.govt.nz/assets/About-us/Future-State/2017-trials-and-pilots/2017-Trials-Pilots-Evaluation-Report-Publication.pdf

New Zealand Qualifications Authority (NZQA). (2017). *National Certificate of Educational Assessment (NCEA) external digital assessment 2016 trials and pilots: User experience evaluation report.* Retrieved 30th October, 2018 from https://www.nzqa.govt.nz/assets/About-us/Future-State/2016-trials-and-pilots/2016-Trials-Pilots-User-Experience-Evaluation-Report-Publication-FINAL.pdf

New Zealand Qualifications Authority (NZQA). (2015). *Digital external assessment prototypes project*. Retrieved 30th October, 2018 from https://www.nzqa.govt.nz/assets/About-us/Future-State/DEAP-Summary-Report.pdf

New Zealand Qualifications Authority (NZQA). (2014). *Report on the eMCAT project.* Retrieved October 30th, 2018 from https://www.nzqa.govt.nz/assets/About-us/Our-role/innovation/2014-eMCAT-report-final.pdf

O'Leary, M., Scully, D., Karakolidis, A., & Pitsia, V. (2018). The state-of-the-art in digital technology-based assessment. *European Journal of Education, 53*(2), 160–175. https://doi.org/10.1111/ejed.12271

Öqvist, M., & Nouri, J. (2018). Coding by hand or on the computer? Evaluating the effect of assessment mode on performance of students learning programming. *Journal of*

Computers in Education, *5*(2), 199–219. https://doi.org/10.1007/s40692-018-0103-3

Organisation for Economic Co-Operation and Development (OECD). (2017). *PISA 2015 Technical Report: Test design and development*. Paris: OECD Publishing. Retrieved April 12th 2018 from https://www.oecd.org/pisa/sitedocument/PISA-2015-Technical-Report-Chapter-2-Test-Design-and-Development.pdf

Paek, P. (2005). *Recent Trends in Comparability Studies. Pearson Educational Measurement Research Report 05-05*. Retrieved September 9th, 2018 from https://pdfs.semanticscholar.org/1379/35ac01dd3822065544eb74c1e9458da921f5.pdf

Parshall, C. & Harmes, J. (2008). The design of innovative item types: Targeting constructs, selecting innovations, and refining prototypes. *CLEAR Exam Review, 19*(2), 18–25. Retrieved September 12th, 2018 from https://www.clearhq.org/resources/CLEAR _summer08_4.pdf#page=20

Parshall, C. G., Spray, J., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Parsons, D. & Haden, P. (2006). Parsons' programming puzzles: a fun and effective learning tool for first programming courses. In *Proceedings of the 8th Australasian Conference on Computing Education: Conferences in Research and Practice in Information Technology Series* (pp. 157-163). ACE: Australia. Retrieved October 25th, 2018 from http://delivery.acm.org/10.1145/1160000/1151890/p157-parsons.pdf?ip=136.206.193.128&id=1151890&acc=PUBLIC&key=846C3111CE4A4710%2E821500BF45340188%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1544434463_3e9f0b25e8d6f312183a86e0a9da0e41

Partnership for Assessment of Readiness for College and Careers (PARCC). (2018a). *PARCC Homepage* [website]. Retrieved October 8th, 2018 from https://parcc.pearson.com/

Partnership for Assessment of Readiness for College and Careers (PARCC). (2018b). *Infrastructure Trial: Readiness Guide Version 7.0*. Retrieved October 8th, 2018 from https://parcc.pearson.com/

Partnership for Assessment of Readiness for College and Careers (PARCC). (2018c). *PARCC Practice Tests* [website]. Retrieved October 8th, 2018 from https://parcc.pearson.com/practice-tests/

Partnership for Assessment of Readiness for College and Careers (PARCC). (2014). *PARCC field test: Lessons learned*. Retrieved October 20th, 2018 from https://parcc-assessment.org/content/uploads/2014/09/field-test-lessons-learned-final_0.pdf

Piech, C., & Gregg, C. (2018). BlueBook: A computerized replacement for paper tests in Computer Science. In *Proceedings of the 49th Technical Symposium on Computer Science Education, SICCSE, 2018*. (pp. 562-567). https://doi.org/10.1145/3159450.3159587

Plimmer, B. & Apperley, M. (2007). Making paperless work. *In Proceedings of the 8th ACM SIGCHI New Zealand Chapter's International Conference on Computer-Human Interaction: Design Centered HCI.* (pp. 1–8). ACM: New York. doi: 10.1145/1278960.1278961

Preece, J., Rogers, Y., Benyon, D., Holland, S., & Carey, T. (1994). H*uman computer interaction*. Wokingham: Addison-Wesley.

Prisacari, A.A., & Danielson, J. (2017). Computer-based versus paper-based testing: Investigating testing mode with cognitive load and scratch paper use. *Computers in Human Behavior*, *77*, 1–10. https://doi.org/10.1016/j.chb.2017.07.044

Radio New Zealand. (February 2018). N*ZQA falsely fails students taking digital exams* [article]. Retrieved 29th November, 2018 from https://www.radionz.co.nz/news/national/ 349917/nzqa-falsely-fails-students-taking-digital-exams

Rajala, T., Kaila, E., Lindén, R., Kurvinen, E., Lokkila, E., Laakso, M.J., & Salakoski, T. (2016). Automatically assessed electronic exams in programming courses. In *Proceedings of the Australasian Computer Science Week Multiconference, ACSW '16*. (pp. 1–8). New York,: ACM Press. https://doi.org/10.1145/2843043.2843062

Rogers, W.T., & Bateson, D.J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. A*pplied Measurement in Education, 4*, 159–183.

Russell, M. (2016). A framework for examining the utility of technology-enhanced items. *Journal of Applied Testing Technology*, *17*(1), 20–32. Retrieved from http://www.jattjournal.com/index.php/atp/article/view/89189/67798%5Cnhttp://www.jattjournal.com/

Russell, M., & Tao, W. (2004). Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research & Evaluation, 9*(1). Retrieved from http://pareonline.net/getvn.asp?v=9&n=1

Sanchez, C. A., & Goolsbee, J.Z. (2010). Character size and reading to remember from small displays. *Computers and Education, 55(3)*, 1056–1062. doi: 10.1016/j.compedu.2010.05.001.

Sarkar, A. (2015). The impact of syntax colouring on program comprehension. In *Proceedings of the 26th Annual Conference of the Psychology of Programming Interest*

Group (PPIG 2015), (pp. 1-9). Bournemouth: PPIG. Retrieved November 28th, 2018 from http://www.ppig.org/sites/default/files/2015-PPIG-26th-Sarkar.pdf

Sarpong, K., Arthur, J., & Amoako, P. (2013). Causes of failure of students in computer programming courses: The teacher-learner perspective. *International Journal of Computer Applications, 77*(12), 27-32. doi: 10.5120/13448-1311

Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and pencil and paper. *European Journal of Psychological Assessment, 26*, 284–292. https://doi.org/10.1027/1015-5759/a000038.

SchoolLeaver. (November, 2018). *Digital NCEA exams - we're not quite there yet* [article]. Retrieved 29 November, 2018 from http://schoolleaver.nz/news-archive/210-n

Scully, Darina (2017). Constructing Multiple-Choice Items to Measure Higher-Order Thinking. *Practical Assessment, Research & Evaluation, 22*(4). Available online: http://pareonline.net/getvn.asp?v=22&n=4

Shiel, G., Kelleher, C., McKeown, C. & Denner, S. (2016). *Future ready? The performance of 15-year-olds in Ireland on science, reading literacy and mathematics in PISA 2015.* Dublin: Educational Research Centre.

Shuhidan, S., Hamilton, M. & D'Souza, D., (2010) Instructor perspectives of multiple-choice questions in summative assessment for novice programmers. *Computer Science Education, 20*(3), 229-259. https://doi.org/10.1080/08993408.2010.509097

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Steedle, J., McBride, M., Johnson, M. & Keng, L. (2016). Spring 2015 digital devices Comparability study. Retrieved March 8th, 2018 from http://blogs.edweek.org/edweek/DigitalEducation/PARCC%20Device%20Comparability%202015%20%28first%20operational%20year%29_FINAL.PDF

State Examination Commission (SEC). (2018). *Junior Certificate - Mathematics* [exam paper]. Retrieved January 7th from https://www.examinations.ie/tmp/1546856873_7413358.pdf

Strain-Seymour, E., Craft, J., Davis, 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 November 7th, 2017 from http://researchnetwork.pearson.com/wp-content/uploads/Testing-on-Tablets-PartI.pdf

Thompson, E., Luxton-Reilly, A., Whalley, J., Hu, M., & Robbins, P. (2008). Bloom's taxonomy for CS assessment. In *Proceedings of the tenth conference on Australasian computing education*. Sydney: Australian Computer Society, Inc.

Tidwell, J. (2010). Designing Interfaces (2nd Edition). O'Reilly Media: Canada. Retrieved October 25th 2018 from http://internativa.com.br/mobile/Livro%20-%20Designing%20Interfaces,%202nd%20Edition,%202010.pdf

U.S. Department of Education. (2015). *Peer review of State assessment systems, non-regulatory guidance for states. September 25, 2015*. Washington: USED. Retrieved May 2nd, 2018, from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf.

Villar, A., Callegro, M. & Yang, Y. (2013). Where am I? A meta-analysis of experiments on the effects of progress indicators for web surveys. *Social Science Computer Review, 31(6)*, 744-762. DOI: 10.1177/0894439313497468.

Vispoel, W. (2000). Reviewing and Changing Answers on Computerized Fixed-Item Vocabulary Tests. *Educational and Psychological Measurement, 60* (3), 371–384

Vorstenbosch, M., Bouter, S., van den Hurk, M., Kooloos, J., Bolhuis, S., & Laan, R. (2014). Exploring the validity of assessment in anatomy: Do images influence cognitive processes used in answering extended matching questions? *Anatomical Sciences Education*, *7*(2), 107–116. https://doi.org/10.1002/ase.1382

Walker, R. & Handley, Z. (2016). Designing for learner engagement with computer-based testing. *Research in Learning Technology, 24*, 1-14. doi: http://dx.doi.org/10.3402/rlt.v24.30083.

Wan, L., & Henly, G. A. (2012). Measurement Properties of Two Innovative Item Formats in a Computer-Based Test. *Applied Measurement in Education*, *25*(1), 58–78. https://doi.org/10.1080/08957347.2012.635507org/10.1080/08957347.2012.635507

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.

Way, W.D., Davis, L., Keng, L. & Strain-Seymour, E. (2015). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology in testing: Improving educational and psychological measurement* (pp. 260-284). UK: Routledge.

Wertheimer M. (1923). Untersuchungen zur Lehre von der Gestalt, II [Laws of organization in perceptual forms]. *Psychologische Forschung, 4*, 301–350. In W. D. Ellis (Ed.), (1938). A source book of Gestalt psychology (pp. 71–94). Routledge & Kegan Paul Ltd: London.

Winslow, L.E. (1996). Programming pedagogy - A psychological overview. *ACM SIGCSE Bulletin, 28*(3), 17-22. Retrieved December 4th, 2018 from http://delivery.acm.org/10.1145/240000/234872/p17-winslow.pdf?ip=136.206.237.150&id=234872&acc=ACTIVE%20SERVICE&key=84

6C3111CE4A4710%2E821500BF45340188%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1543926875_3c2d9942ac09b77cc79e360cb0f57bee

Winter, P. (2010). Comparability and test variations. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 1-11). Washington, DC: Council of Chief State School Officers. Retrieved 3rd March, 2018 from https://files.eric.ed.gov/fulltext/ED543067.pdf.

Woo, A., Kim, D., & Qian, H. (2014). *Exploring the psychometric properties of innovative items in CAT. National Council of State Boards of Nursing*. Paper presented at 2014 MARCES Conference: Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring from an Interdisciplinary Perspective. University of Maryland. Retrieved from https://marces.org/conference/InnovativeAssessment/5Woo.pdf

Woodford, K. & Bancroft, P. (2005). Multiple choice questions not considered harmful. I*n Proceedings of the 7th Australasian Conference on Computing Education (ACE2005)*, p. 109-115. Retrieved December 1st, 2018 from http://crpit.com/confpapers/CRPITV42Woodford.pdf.

Appendix 1

Online NCEA Exam Checklist (NZQA, 2018b)

Online NCEA Exam Checklist for Candidates

NEW ZEALAND QUALIFICATIONS AUTHORITY
MANA TOHU MĀTAURANGA O AOTEAROA

QUALIFY FOR THE FUTURE WORLD
KIA NOHO TAKATŪ KI TŌ ĀMUA AO!

Candidates participating in NCEA Digital Pilots for the following subjects are strongly advised to complete the items on this checklist:

- Classical Studies, Level 1-3
- English, Level 1-3
- Media Studies, Level 1-3

To PREPARE for your exam

Visit www.nzqa.govt.nz/trials-pilots and complete the following:

	Tick off when done
View the Help videos on the NZQA website.	
Login and complete the Familiarisation Activity Tool.	
Complete the previous years' digital exam in your subject. Links to these are also found in the Familiarisation activity section.	

NZQA will provide information to your school about whether you have completed the Familiarisation activities.

CHECK your device

If you are using your personal device check that it:

Is fully charged.	
Has notifications, screensavers and automatic updates turned off.	
Has screen resolution is set to a minimum of 600*800.	
Is virus-free.	
Has one of the following as the default browser: ○ Chrome (v54 or higher) ○ Firefox (v49 or higher) ○ Safari (v10 or higher) We recommend you update your browser to the most recent version	

What you need on the day:

- Your NCEA Candidate Admission slip - the USERNAME and PASSWORD for your online NCEA exam is printed on your slip.
    - Username = NSN
    - Password – printed directly below your NSN
- Your personal device, fully charged, and power cable – if the school is not providing a device.
- Pens, pencils, etc (contained in a clear sealed bag) in case you need to switch to paper.

What to expect in the exam:

If you have completed a Digital Trial, previous digital exam or the Familiarisation Activity Tool (highly recommended) most aspects of the online NCEA exam will be familiar, however there are a few key things to remember:

If you	**What will happen**
Cannot login	Let the supervisor know, they will assist you and contact technical support if required.
Leave the examination window (screen) e.g. try to access something else on your device or on the internet	You will receive a WARNING and may be LOCKED OUT of the exam. The supervisor can unlock the examination but will report the breach to NZQA. Note: this is different from what you may have experienced in a Trial.
Lose internet connection	A WARNING message will show on the supervisor's screen. They will contact technical support if required.
Have a problem with the online exam, or your device, which can't be resolved quickly	You can switch to paper. You will be given extra time if required. The supervisor will determine how much.
Decide not to do your exam online	You can switch to paper. If you have started a standard online, you must: • complete the standard digitally OR • copy your responses for that standard into the booklet. You will not be given extra time for this.

For more information:

Visit www.nzqa.govt.nz/trials-pilots for access to useful links, including the Student Factsheet and Familiarisation activities.